

STATGRAPHICS® Centurion XVI

Benutzerhandbuch

STATGRAPHICS® CENTURION XVI BENUTZERHANDBUCH

---

www.STATGRAPHICS.com

Referenz: *STATGRAPHICS® Centurion XVI Benutzerhandbuch*

Gedruckt in den Vereinigten Staaten von Amerika, USA.

Inhalt

Vorwort

Dieses Handbuch hat zum Ziel, die Anwender von STATGRAPHICS Centurion XVI in die grundlegende Handhabung des Programms und seine vielfältigen Anwendungsmöglichkeiten bei der Analyse von Daten einzuführen. Es bietet einen umfassenden Überblick über das System, einschließlich der Installation, dem Daten-Management , dem Erstellen von statistischen Analysen sowie den Optionen für Druck und Publikation. Ein Anliegen bei der Gestaltung des Handbuchs war, den Anwendern ein Hilfmittel an die Hand zu geben, mit dem sie schnell und ohne lange Vorlaufszeit eigenständig Analysen entwerfen und durchführen können. Daher konzentriert sich sein Inhalt mehr auf die wichtigsten Funktionalitäten des Programms als darauf, bis in jedes Detail zu gehen. Für detaillierte Informationen steht den Anwendern das Hilfe-System von STATGRAPHICS Centurion XVI zur Verfügung. Hier erhalten Sie nicht nur eine große Bandbreite an Zusatzinformationen, sondern haben auch Zugriff auf die Erläuterungen zu jeder der ungefähr 160 enthaltenen statistischen Prozeduren, die als separate PDF-Dokumente aufgerufen werden können.

Die ersten neun Kapitel dieses Handbuchs befassen sich mit den grundsätzlichen Einsatzmöglichkeiten des Programms. Zwar ist es durchaus möglich, vieles davon allein durch die praktische Anwendung des Programms zu erlernen, doch ermöglicht ein sorgfältiges Durcharbeiten dieser Kapitel, die Anwendungen schnell, fehlerfrei und vollständig auszuführen.

Die anschließenden sieben Kapitel verstehen sich als sogenannte Tutorials (Übungen) zur

1. Einführung in die am häufigsten gebrauchten statistischen Analysen

2. Veranschaulichung, wie die einzigartigen Funktionalitäten von STATGRAPHICS Centurion XVI den Prozess der Datenanalyse vereinfachen.

Bitte nutzen Sie die Tutorials, um herauszufinden, wie Sie Ihre aktuellen Daten mit STATGRAPHICS Centurion XVI optimal analysieren können.

HINWEIS: Das Handbuch ist im *Hilfe*-Menü als PDF-Dokument mit farbigen Abbildungen enthalten. Die Datendateien und StatFolios, auf die hier Bezug genommen wird, sind im Programm enthalten.

StatPoint Technologies, Inc.
August, 2009

Kapitel

1

Einführung

Installieren von STATGRAPHICS Centurion XVI, Inbetriebnahme des Programms und Erstellen einer einfachen Datendatei.

1.1 Installation

STATGRAPHICS Centurion XVI ist auf zwei Wegen erhältlich: zum einen kann das Programm in Form einer einzigen Datei aus dem Internet auf Ihren Computer geladen werden, zum anderen als CD-ROM mit mehreren Dateien. Um das Programm zu starten, muss die Software zunächst auf Ihrer Festplatte installiert werden. Wie bei den meisten Windows-Anwendungen ist die Installation sehr einfach:

Schritt 1: Wenn Sie das Programm auf einer CD erhalten haben, legen Sie diese in das CD-ROM-Laufwerk Ihres Computers ein. Nach einigen Augenblicken öffnet sich automatisch ein Setup-Programm. Sollte dies nicht der Fall sein, öffnen Sie bitte den Windows Explorer und führen die Datei **gcinstall.exe** aus dem Hauptverzeichnis der CD aus.

Wenn Sie das Programm über das Internet geladen haben, wählen Sie die geladene Datei an ihrem entsprechenden Speicherort per Doppelklick aus, um die Installation zu starten.

Schritt 2: Es wird Ihnen nun eine Reihe von Dialogboxen angezeigt. Wenn Sie das Programm von einer CD aus starten, werden Sie in der ersten Dialogbox gebeten, die zu installierende(n) Sprache(n) auszuwählen:

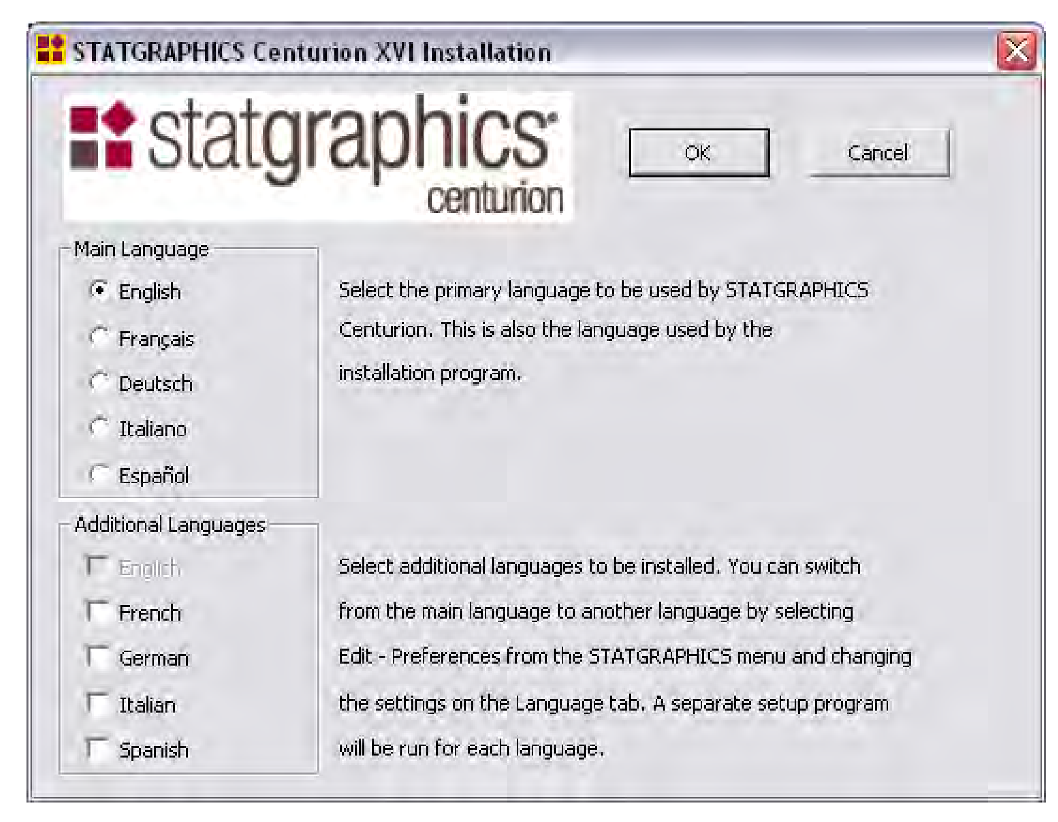


Abbildung 1.1: Dialogbox für die Sprachauswahl

Bitte wählen Sie die gewünschte Haupt-Programmsprache sowie eine oder mehrere zusätzliche Sprachen aus. Die Hauptsprache wird dann während der Installation und als Standardsprache beim ersten Programmstart verwendet. Wenn Sie zusätzliche Sprachen installieren, können Sie nachher im Programm zwischen den Sprachen wechseln. Gehen Sie dazu in das Hauptmenü zum Menüunterpunkt *Bearbeiten – Voreinstellungen*.

Wenn Sie das Programm per Download aus dem Internet installieren, muss das Setup für jede gewünschte Sprache separat durchgeführt werden.

> ANMERKUNG: Während der Testphase können Sie auf jede in STATGRAPHICS Centurion XVI enthaltene Sprach zugreifen. Bei der Lizenzierung muss dann jedoch angegeben werden, welche Sprache als Hauptsprache und welche Sprache(n) als zusätzliche Sprache(n) gewünscht werden. Nach der Testphase haben Sie dann ausschließlich auf die hier spezifizierten Sprachen Zugriff.

Schritt 3: STATGRAPHICS Centurion XVI verwendet InstallShield, um das Programm auf Ihrem Computer zu installieren. Der InstallShield-Assistent leitet Sie mit Hilfe einer Reihe von Dialogboxen durch den Installationsprozess. Die erste Dialogbox begrüßt Sie zu STATGRAPHICS Centurion XVI. Hier wählen Sie einfach die Schaltfläche *Weiter* aus.



Abbildung 1.2: Willkommen-Dialogbox

ANMERKUNG: Sie benötigen Administratoren-Rechte, um STATGRAPHICS Centurion XVI auf Ihrem Computer zu installieren und zu aktivieren. Während der Installation sollte Ihnen daher ein Systemadministrator zur Seite stehen. Wir empfehlen Ihnen, das Programm nur zu **installieren und zu aktivieren**, wenn ein Administrator Sie unterstützen kann.

Schritt 4: Die zweite Dialogbox enthält die Lizenzvereinbarung für die Software:



Abbildung 1.3: Dialogbox mit der Lizenzvereinbarung

Bitte lesen Sie die Lizenzvereinbarung sorgfältig. Wenn Sie die Lizenzbedingungen akzeptieren, markieren Sie bitte das entsprechende Optionsfeld. Wählen Sie dann *Weiter*, um fortzufahren. Wenn Sie den Lizenzbedingungen nicht zustimmen, wählen Sie *Abbrechen*. In diesem Fall können Sie das Programm nicht starten.

Schritt 5: In der nächsten Dialogbox geben Sie bitte Informationen zu Ihrer Person bzw. der Person, die das Programm nutzen wird, ein:



Abbildung 1.4: Dialogbox mit den Benutzerinformationen

Bitte geben Sie alle benötigten Informationen ein. Sie können das erste Optionsfeld markieren, wenn Sie allen Personen, die Ihren Computer nutzen, den Zugriff auf STATGRAPHICS Centurion XVI erlauben möchten.

Schritt 6: Die nächste Dialogbox zeigt an, in welches Verzeichnis das Programm installiert wird:

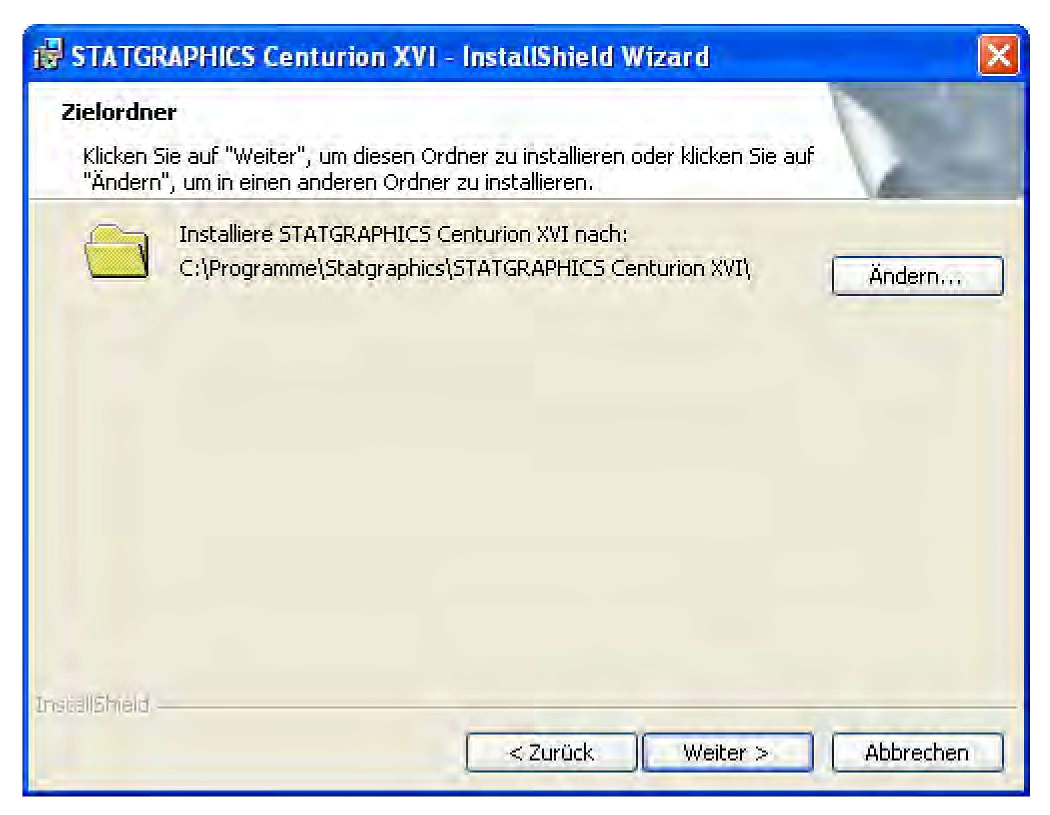


Abbildung 1.5: Dialogbox mit dem Zielverzeichnis der Installation

STATGRAPHICS Centurion XVI wird standardmäßig in ein eigenes Unterverzeichnis von *Programme* mit der Bezeichnung *Statgraphics – STATGRAPHICS Centurion XVI* installiert. Wenn Sie das Programm auf einem Netzwerk-Server installieren, wählen Sie als Speicherziel ein Verzeichnis aus, für das alle potenziellen Anwender eine Leseberechtigung haben. Schreibrechte sind hierbei nicht erforderlich. Eine vollständige Anleitung für die Installation des Programms in einem Netzwerk ist in der Textdatei *Liesmich.txt* enthalten, die Sie auf der Installations-CD bzw. bei den aus dem Internet geladenen Installationsdateien finden werden.

Schritt 7: Mit der nächsten Dialogbox können Sie den gewünschten Installationsumfang bestimmen:



Abbildung 1.6: Dialogbox zur Auswahl des Installationsumfangs

Wählen Sie eine der folgenden Möglichkeiten aus:

Standard – installiert das Programm, die Hilfedateien, die Dokumentation und die Beispiel-Datendateien. Für dieses Setup werden mindestens 60 MB Speicherplatz auf Ihrer Festplatte benötigt.

Minimal – installiert nur das Programm und die Hilfedateien. Es werden ungefähr 30 MB Speicherplatz auf Ihrer Festplatte benötigt.

Angepasst – installiert ausschließlich die Komponenten, die Sie auswählen.

Mit der Minimal-Installation können Sie Speicherplatz auf Ihrer Festplatte sparen, Sie haben dann jedoch keinen Zugriff auf die Online-Dokumentation und die begleitenden Beispiel-Datendateien.

Schritt 8: Folgen Sie den weiteren Anweisungen, um die Installation abzuschließen. Nach der Installation wird die folgende letzte Dialogbox angezeigt:

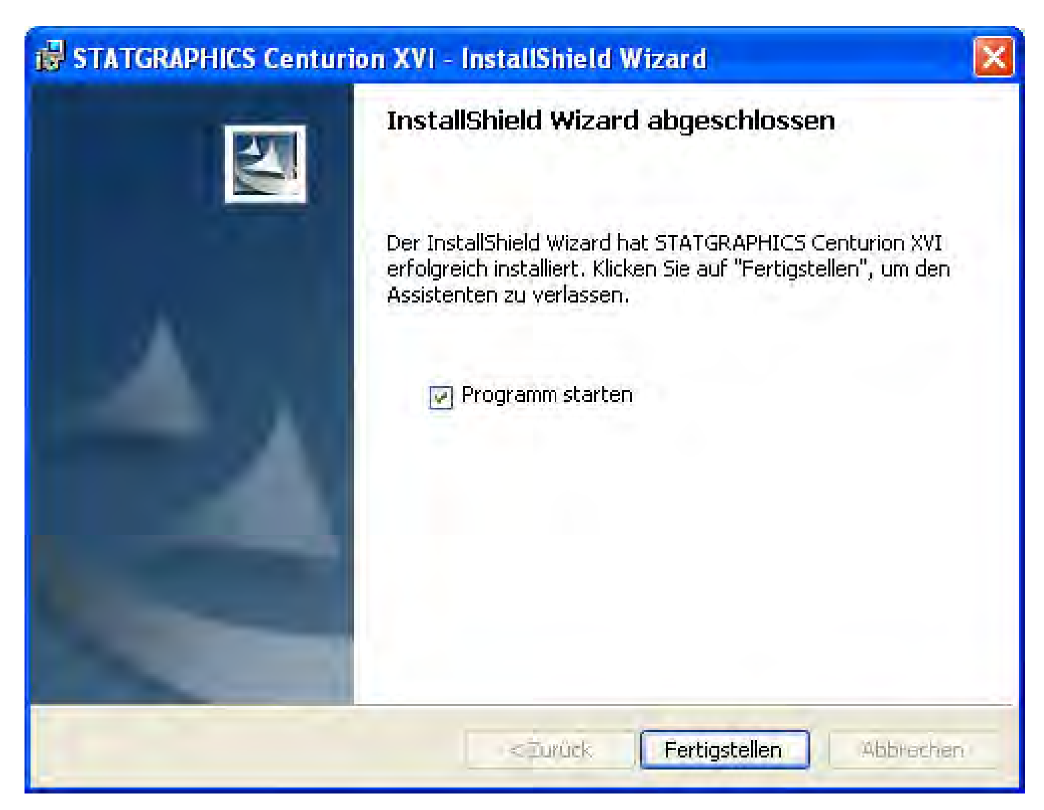


Abbildung 1.7: Dialogbox zum Abschließen der Installation

Wählen Sie *Fertigstellen*, um die Installation abzuschließen. .Sie können das Kontrollkästchen *Programm starten* markieren, um STATGRAPHICS Centurion XVI direkt zu starten. Oder Sie folgen den unten stehenden Anleitungen.

1.2 Inbetriebnahme des Programms

Während der Installation fügt STATGRAPHICS Centurion XVI eine Verknüpfung zu dem Programm dem Windows *Start*-Menü und dem Desktop Ihres Computers hinzu. Folgen Sie diesen Anweisungen, um das Programm erstmalig zu starten:

Schritt 1: Wählen Sie die Verknüpfung auf Ihrem Desktop mit der linken Maustaste an oder öffnen Sie das Windows *Start*-Menü in der linken unteren Ecke Ihres Bildschirms und wählen

Sie hier unter *Programme – STATGRAPHICS Centurion XVI* das *Statgraphics*-Symbol aus. Über den Windows Explorer finden Sie das Programm zudem in dem Verzeichnis *Programme – Statgraphics – STATGRAPHICS Centurion XVI*; hier wählen Sie das Anwendungssymbol *sgwin*, um das Programm auszuführen.

Schritt 2: Während STATGRAPHICS Centurion XVI geladen wird, öffnet sich ein neues Fenster. Beim ersten Programmaufruf begrüßt Sie eine *Willkommen*-Dialogbox:



Abbildung 1.8: Willkommen-Dialogbox

Bitte wählen Sie eine der zwei angezeigten Optionen:

1. *Bewerten*, um das Programm 30 Tage kostenfrei zu testen, bevor Sie es käuflich erwerben.

2. *Aktivieren*, wenn Sie das Programm bereits käuflich erworben und Ihre Seriennummer erhalten haben.

Wenn Sie *Bewerten* gewählt haben, wird folgende Dialogbox angezeigt:



Abbildung 1.9: Dialogbox für den Bewerten-Modus

In der Dialogbox wird Ihnen ein 16-stelliger Produktschlüssel angezeigt. Um das Programm im Testmodus aktivieren zu können, benötigen Sie einen *Aktivierungskode*. Sie erhalten Ihren Aktivierungskode, indem Sie eine der unter *Schritt 2* beschriebenen Schaltflächen anklicken:

1. Die Schaltfläche *1. Drücken Sie hier* versendet automatisch über das Internet eine Nachricht an StatPoint Technologies mit der Anforderung eines Aktivierungskodes. Per Web Service wird Ihnen daraufhin Ihr persönlicher Kode an die vorher eingegebene E-Mail-Adresse gesendet.

2. Die Schaltfläche *2. Drücken Sie hier* öffnet Ihr Standard-E-Mail-Programm und fügt alle benötigten Informationen in eine neue E-Mail ein, die Sie dann direkt an StatPoint versenden können. Solche E-Mail-Anfragen können allerdings nur während der normalen Geschäftszeiten (USA) bearbeitet werden.

Um Verzögerungen zu vermeiden, verwenden Sie bitte, wenn möglich, die erste Schaltfläche.

> ANMERKUNG: Benutzer, die eine akademische Site Lizenz aktivieren, müssen die erste Installations-Methode wählen. Der Aktivierungskode wird nur an die beim Kauf registrierte E-Mail-Adresse versendet und sollte dann intern an die jeweiligen User verteilt werden.

Schritt 3: Nach Bearbeitung Ihrer Anforderung erhalten Sie Ihren Aktivierungskode per E-Mail. Bitte geben Sie diesen Kode in das unter *Schritt 3* vorgegebene Feld ein und klicken Sie die Schaltfläche *Aktivieren* an. Wenn der Kode und der Produktschlüssel zusammen passen, wird Ihnen folgende Nachricht angezeigt:



Abbildung 1.10: Aktivierungs-Nachricht

Klicken Sie *OK,* um in das Programm zu gelangen.

> ANMERKUNG 1: Unter Microsoft Vista oder Windows 7 wird das gewohnte Doppelklicken auf das STATGRAPHICS-Symbol auf Ihrem Desktop möglicherweise nicht korrekt funktionieren. In diesem Fall klicken Sie bitte mit der rechten Maustaste

auf das Symbol und wählen in dem nun sichtbaren Pop-Up-Menü die Option *Als Administrator ausführen* aus.

ANMERKUNG 2: Wenn Sie STATGRAPHICS Centurion später einmal auf einem anderen Computer installieren, werden Sie den Aktivierungs-Prozess noch einmal durchführen müssen, da das Programm für jeden Computer einen eigenen Produktschlüssel erzeugt.

Schritt 4: Beim erstmaligen Aufrufen des Programms werden Sie gefragt, welches Menüsystem Sie verwenden möchten:

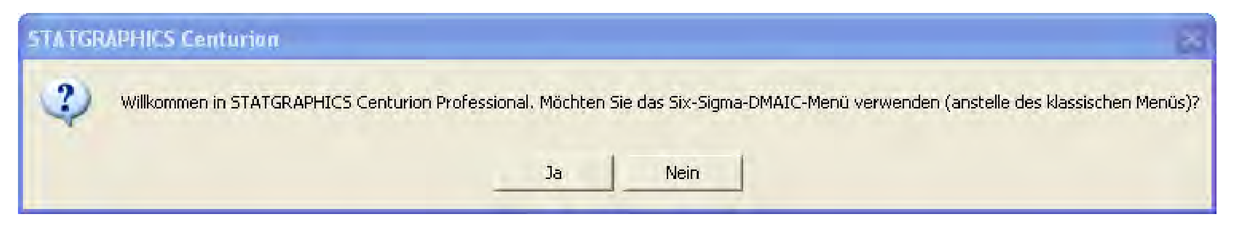


Abbildung 1.11: Menüauswahl-Fenster

Sie können nun auswählen zwischen dem klassischen STATGRAPHICS-Menü, das die statistischen Prozeduren in die Themenbereiche *Diagramme, Beschreiben, Vergleichen, Beziehungen, Vorhersagen, SPC* und *DOE* gliedert, und dem Six-Sigma-Menü, das die Prozeduren unter *Definieren, Messen, Analysieren, Verbessern, Regeln* und *Vorhersagen* anordnet. Beide Menüs beeinhalten dieselben Prozeduren, sie unterscheiden sich jedoch in deren Anordnung. Die Auswahl des Menüsystems können Sie später jederzeit ändern, indem Sie im *Bearbeiten*-Menü unter *Voreinstellungen* die gewünschte Option auswählen.

Schritt 5: Nun wird das STATGRAPHICS Centurion XVI-Anwendungsfenster geöffnet:

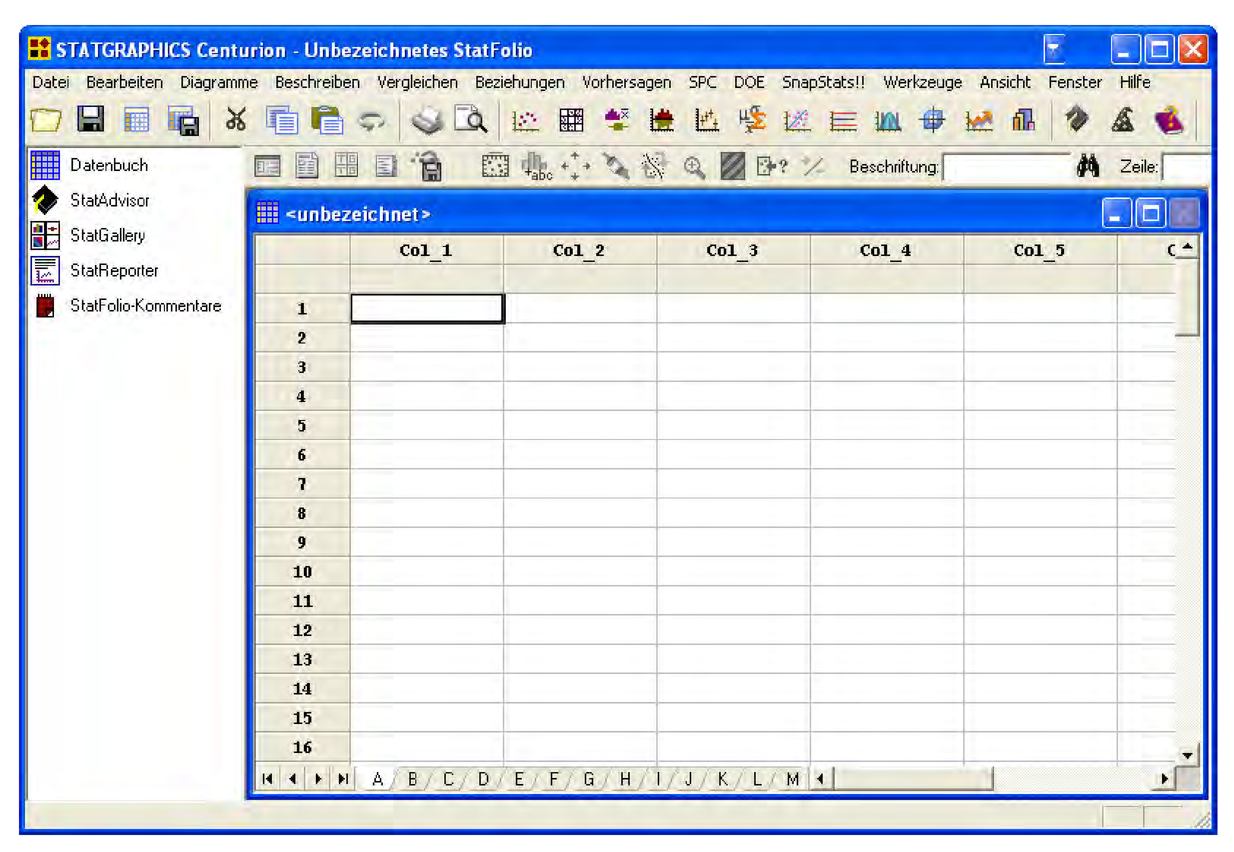


Abbildung 1.12: Haupt-Programmfenster von STATGRAPHICS Centurin XVI

In den nachfolgenden Abschnitten beschreiben wir beispielhaft, wie Sie eine Datendatei aus den Daten einer Volkszählung der USA des Jahres 2000 erstellen.

1.3 Eingeben der Daten

Um Daten in STATGRAPHICS Centurion XVI zu analysieren, werden diese in ein STATGRAPHICS-*Datenbuch* eingegeben. Ein Datenbuch besteht aus 26 *Datenblättern*, die mit den Buchstaben *A* bis *Z* bezeichnet sind und jeweils eine Matrix aus Zeilen und Spalten enthalten:

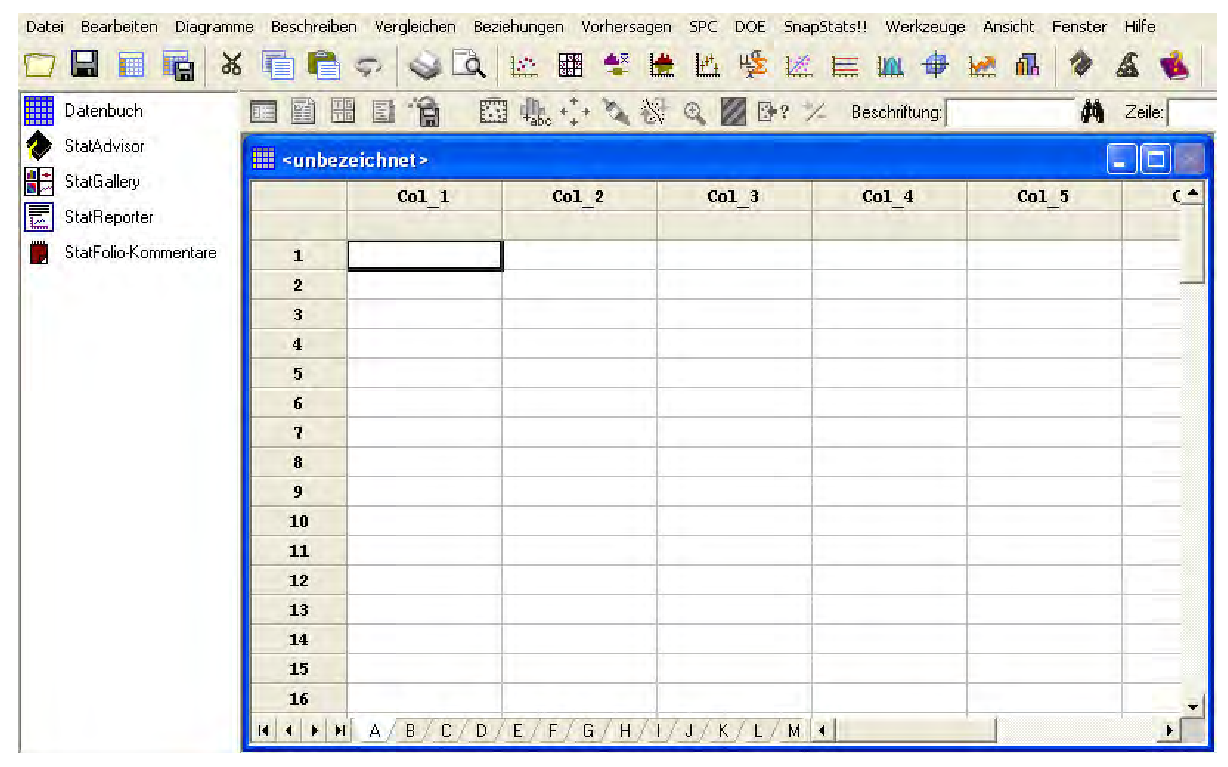


Abbildung 1.13: Das STATGRAPHICS-Datenbuch

In einem typischen Datenblatt enthält jede Zeile Informationen zu einer einzelnen Probe, Fall oder Beobachtung, während jede Spalte eine Variable repräsentiert.

Nehmen Sie z. B. an, Sie möchten mit STATGRAPHICS Centurion XVI die Daten der US-amerikanischen Bevölkerungszählung analysieren. Hier ein Auszug aus den Daten:

Staat	Bevölkerung	Durchschnittsalter	% Frauen	Pro Kopf Einkommen
Alabama	4.447.100	35,8	51,7	$18.819
Alaska	626.932	32,4	48,3	$22.660
Arizona	5.130.632	34,2	50,1	$20.275
Arkansas	2.673.400	36,0	51,2	$16.904
California	33.871.648	33,3	50,2	$22.711
Colorado	4.301.261	34,3	49,6	$24.049

Abbildung 1.14: Auszug aus den Daten der US-amerikanischen Bevölkerungszählung des Jahres 2000

Wenn Sie die Daten in ein STATGRAPHICS Centurio XVI-Datenblatt eingeben, werden die Informationen zu jedem einzelnen Staat jeweils in eine eigene Zeile eingefügt. Insgesamt werden fünf Spalten erstellt, in welche die Namen der Staaten sowie die Daten der Zählung eingetragen werden.

Um derartige Daten in STATGRAPHICS Centurion XVI einzugeben, können Sie zwischen den folgenden Möglichkeiten wählen:

1. Tippen Sie die Daten direkt in das STATGRAPHICS Centurion XVI-Datenblatt ein.

2. Geben Sie die Daten in einem anderen Programm wie z. B. Excel ein und kopieren Sie sie anschließend nach STATGRAPHICS Centurion XVI bzw. lassen Sie das Programm die Daten einlesen.

In diesem Abschnitt wählen wir die erste Möglichkeit. Um zu beginnen klicken Sie doppelt auf die Kopfzeilen der ersten Spalte mit dem Titel *Col_1*. Sie öffnen damit eine Dialogbox, mit der Sie die Eigenschaften der Spalte einstellen bzw. ändern können.



Abbildung 1.15: Dialogbox zur Definition der Inhalte der Spalten

In einem STATGRAPHICS Centurion XVI-Datenblatt wird jeder Spalte ein Name, ein Kommentar sowie der einzugebende Datentyp zugeordnet:

- *Name* – Geben Sie jeder Spalte einen eindeutigen Namen aus 1 bis 32 Zeichen. Das Programm benutzt die Namen, um die für eine gewählte statistische Prozedur zu analysierenden Variablen zu bestimmen. Zudem dienen die Namen zur Standard-Bezeichnung der Diagrammachsen. Die Namen können alle Zeichen, auch Leerzeichen, enthalten; zwischen Groß- und Kleinbuchstaben wird nicht unterschieden. Wenn Sie den gleichen Namen für mehr als eine Spalte in ein und demselben Datenblatts verwenden, erhalten Sie eine Fehlermeldung.

- *Kommentar* – Hier können Sie den Inhalt der Spalte beschreiben (bis zu 64 Zeichen). Kommentare sind optional. Vorhandene Kommentare werden in der zweiten Zeile der Kopfzeilen dargestellt.

- *Typ* – Legen Sie den Typ der Daten fest, die in die Spalte eingegeben werden. In unserem Beispiel wird der Typ *Text* für die Staatennamen gewählt. Die anderen Spalten können als *Numerisch* (Standardeinstellung) belassen oder als *Integer* oder *Festkomma* (wenn Sie die

Datentypen einschränken möchten) festgelegt werden. Für ausführliche Informationen zu den Variablentypen siehe Kapitel 2.

Wenn Sie eine Spalte definiert haben, wählen Sie *OK*. Bitte definieren Sie fünf Spalten wie folgt:

<unbezeichnet>

	State	Population	Median age	Percent Female	Per Capita Income	Col
	state name		years		average	
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						

A B C D E F G H I J K L M

Abbildung 1.16: STATGRAPHICS Centurion XVI-Datenblatt mit den definierten Spaltennamen

Geben Sie nun, wie bei anderen Spreadsheets auch, die Daten ein, indem Sie die entsprechenden Zellen über die Pfeiltasten Ihrer Tastatur anwählen. Anschließend sollte Ihr Datenblatt folgendermaßen aussehen:

<unbezeichnet>

	State	Population	Median age	Percent Female	Per Capita Income	Col
	state name		years		average	
1	Alabama	4447100	35,8	51,7	18819	
2	Alaska	626932	32,4	48,3	22660	
3	Arizona	5130632	34,2	50,1	20275	
4	Arkansas	2673400	36	51,2	16904	
5	California	33871648	33,3	50,2	22711	
6	Colorado	4301261	34,3	49,6	24049	
7						
8						
9						
10						

A B C D E F G H I J K L M

Abbildung 1.17: STATGRAPHICS Centurion XVI-Datenblatt nach der Eingabe von 6 Daten-Zeilen

Abschließend wird die Datendatei gesichert. Wählen Sie hierzu *Datei – Speichern – Datendatei speichern* im Hauptmenü aus. Geben Sie in der geöffneten Dialogbox den *Dateinamen*, unter dem Sie die Daten sichern möchten, an:

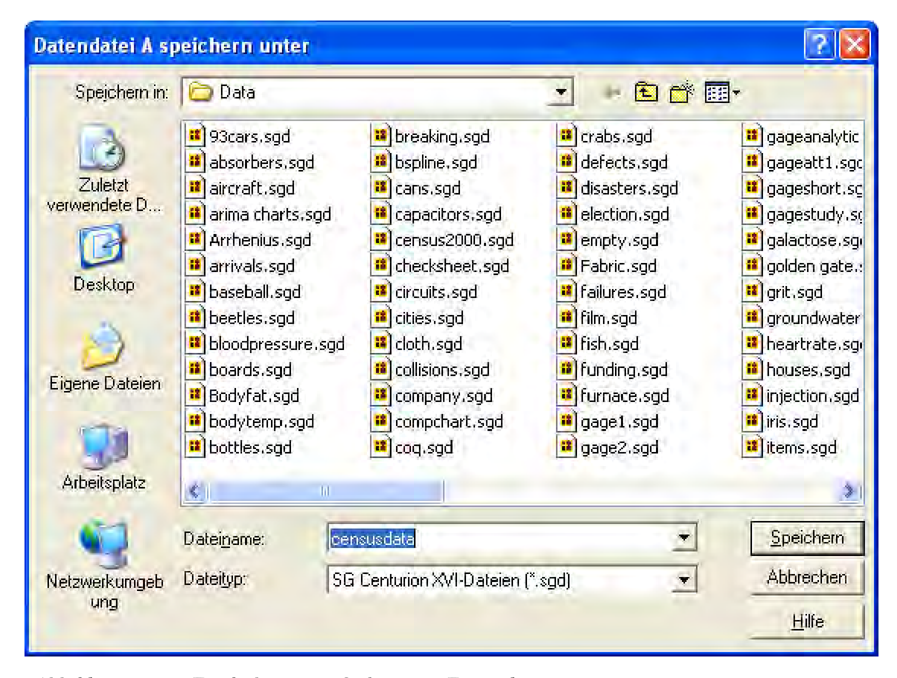


Abbildung 1.18: Dialogbox zum Sichern von Datendateien

In STATGRAPHICS Centurion XVI werden die Datendateien standardmäßig mit der Erweiterung *.sgd*, die die Daten im XML-Format auf der Festplatte ablegt, gespeichert. Beim Speichern der Datei können Sie diese Einstellung über das Feld *Dateityp* ändern und hier den gewünschten Dateityp auswählen.

1.4 Öffnen einer gespeicherten Datendatei

Sobald die Daten in das Datenblatt eingegeben wurden, können sie ausgewertet werden. Um unser Beispiel etwas interessanter zu gestalten, wollen wir die Zensus-Daten aller 50 US-amerikanischen Staaten und des District of Columbia abrufen. Sie sind in STATGRAPHICS Centurion XVI unter dem Dateinamen *census2000.sgd* enthalten. Diese Datendatei öffnen Sie durch Auswählen von *Datei – Öffnen – Datenquelle öffnen* im Hauptmenü. Hier werden Sie gebeten, den Speicherort der gewünschten Datei anzugeben:

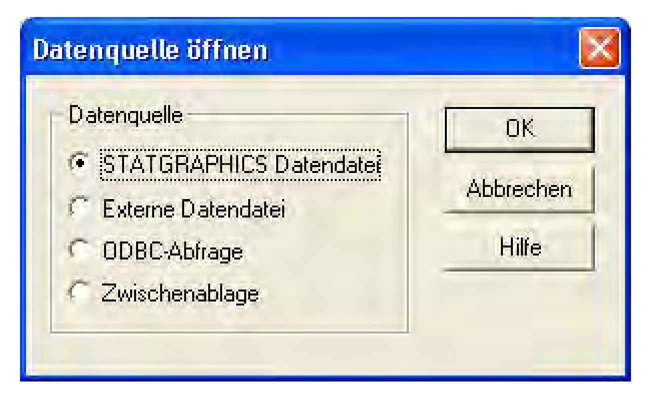


Abbildung 1.19: Dialogbox zur Auswahl der Datenquelle

In unserem Beispiel kann die Voreinstellung *STATGRAPHICS-Datendatei* belassen werden. In der folgenden Dialogbox wählen Sie den Namen der Datei mit den benötigten Daten aus:

Abbildung 1.20: Dialogbox zur Auswahl der zu öffnenden Datendatei

Die Beispieldatei befindet sich im Standardverzeichnis (üblicherweise ist das *C:\Programme \Statgraphics\STATGRAPHICS Centurion XVI\Data*).

Mit dem Öffnen der Datei werden alle 51 Datenzeilen in das Datenblatt geladen:

C:\Programme\Statgraphics\STATGRAPHICS Centurion XVI\Data\census2000.sgd

	State	Population	Median Age	Percent Female	Per Capita Income	Col_
1	Alabama	4447100	35.8	51.7	18819	
2	Alaska	626932	32.4	48.3	22660	
3	Arizona	5130632	34.2	50.1	20275	
4	Arkansas	2673400	36	51.2	16904	
5	California	33871648	33.3	50.2	22711	
6	Colorado	4301261	34.3	49.6	24049	
7	Connecticut	3405565	37.4	51.6	28766	
8	Delaware	783600	36	51.4	23305	
9	D.C.	572059	34.6	52.9	28659	
10	Florida	15982378	38.7	51.2	21557	
11	Georgia	8186453	33.4	50.8	21154	
12	Hawaii	1211537	36.2	49.8	21525	

census2000 B C D E F G H I J K

Abbildung 1.21: Datenblatt mit dem Inhalt der Datei census2000.sgd

1.5 Analysieren der Daten

Sobald die Daten in das STATGRAPHICS Centurion XVI-Datenbuch geladen wurden, kann auf jede der über 160 statistischen Prozeduren wie folgt zugegriffen werden:

1. Durch Auswahl der gewünschten Prozedur aus dem Hauptmenü.

2. Durch Auswahl einer der Schaltflächen auf der Symbolleiste.

3. Durch Aufruf des StatWizard über die Zauberhut-Schaltfläche auf der Symbolleiste.

Wir beginnen unsere Analyse, indem wir die Schwankungen zwischen dem Pro-Kopf-Einkommen der Staaten zusammenfassen. Die am besten geeignete Prozedur, um eine einzige Spalte mit numerischen Daten zusammenzufassen, ist die *Analyse einer Variablen*. Sie berechnet Summenstatistiken wie den Mittelwert und die Standardabweichung. Zudem erstellt sie verschiedene Diagramme, wie z. B. Häufigkeitsdiagramme und Box-Whisker-Plots.

Von wo aus Sie die Prozedur *Analyse einer Variablen* aufrufen, hängt davon ab, welches Menüsystem Sie verwenden:

1. **Klassisches Menü:** wählen Sie *Beschreiben – Numerische Daten – Analyse einer Variablen.*

2. **Six Sigma-Menü:** wählen Sie *Analysieren – Messbare Merkmale – Analyse einer Variablen.*

Wie jede andere statistische Prozedur, beginnt die *Analyse einer Variablen* mit der Anzeige einer Dialogbox, in die bestimmte Daten eingegeben werden müssen:



Abbildung 1.22: Dateneingabe-Dialogbox für die Analyse einer Variablen

Im linken Eingabefeld sind die Namen aller im Datenblatt definierten und Daten enthaltenden Spalten aufgelistet. Um die Daten der Spalte *Pro Kopf Einkommen* zu analysieren, markieren Sie den entsprechenden Namen und fügen ihn durch Anklicken der Schaltfläche mit dem schwarzen Pfeil dem Eingabefeld *Daten* hinzu. Das Feld *Bedingung* lassen Sie frei. (Dieses Feld wird nur benutzt, wenn Sie nur einen Teil der Zeilen auswerten wollen.)

Sobald Sie *OK* auswählen, wird die Dialogbox *Tabellen und Grafiken* geöffnet. Hier wird angezeigt, welche Tabellen und Diagramme für die *Analyse einer Variablen* verfügbar sind. An dieser Stelle wollen wir die Standardauswahl verwenden:

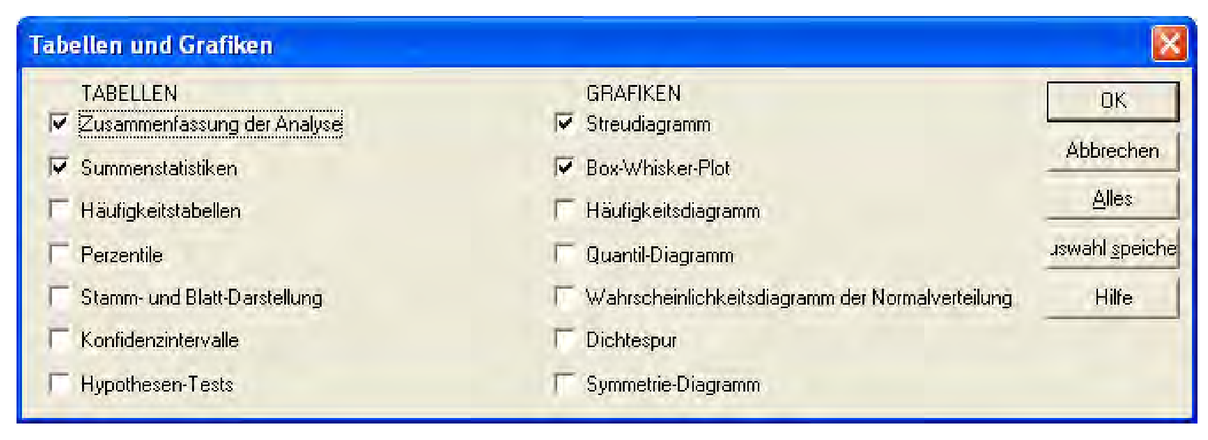


Abbildung 1.23: Dialogbox Tabellen und Grafiken

Wenn Sie auf *OK* klicken, wird ein neues Fenster geöffnet:

Analyse einer Variablen - Per Capita Income

Analyse einer Variablen - Per Capita Income
Datenvariable: Per Capita Income
51 Werte im Bereich von 15853,0 bis 28766,0

Der StatAdvisor
Diese Prozedur ist zur Zusammenfassung einer einzelnen Stichprobe gedacht. Sie k
weiteren werden Konfidenzintervalle berechnet und Hypothesentests durchgeführt
Analyse-Symbolleiste, um auf die verschiedenen Auswerte-Möglichkeiten zugreife

Summenstatistiken für Per Capita Income

Anzahl	51
Arithm. Mittelwert	20934,5
Standardabweichungen	3018,79
Variationskoeffizient	14,4202%
Minimum	15853,0
Maximum	28766,0
Spannweite	12913,0
Stand. Schiefe	1,93465
Stand. Wölbung	0,421446

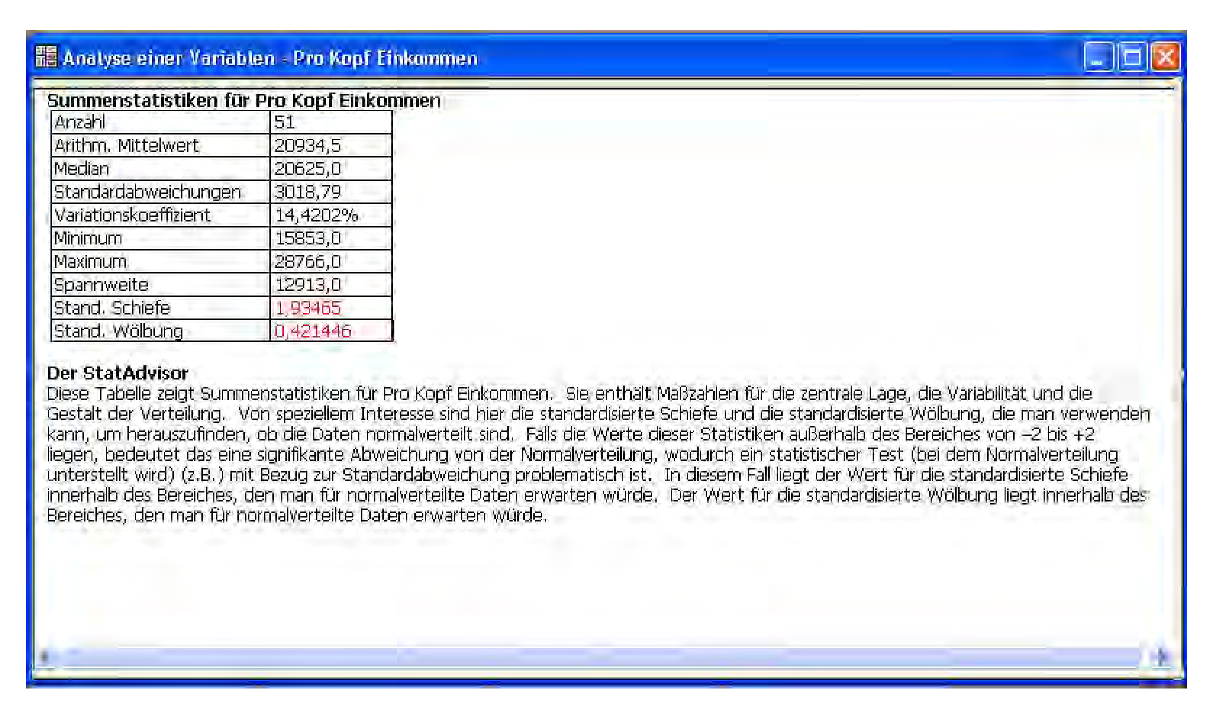
Der StatAdvisor
Diese Tabelle zeigt Summenstatistiken für Per Capita Income. Sie enthält Maß
Verteilung. Von speziellem Interesse sind hier die standardisierte Schiefe und d

Streudiagramm
15 18 21 24 27 30 (X 1000,0)
Per Capita Income

Box-Whisker-Plot
15 18 21 24 27 30 (X 1000,0)
Per Capita Income

Abbildung 1.24: Ergebnisfenster der Analyse einer Variablen

Das Analyse-Fenster enthält mehrere Abschnitte, so genannte *Ergebnisfenster* oder *Panes*, die durch bewegbare Trennbalken unterteilt sind. Die Ergebnisfenster der linken Hälfte beinhalten tabellarische, die Fenster der rechten Hälfte grafische Auswertungen der Daten. Wenn Sie doppelt in das linke untere Fenster klicken, in dem sich die Summenstatistiken befinden, können Sie die Tabelle der Summenstatistiken maximieren:


Analyse einer Variablen - Pro Kopf Einkommen

Summenstatistiken für Pro Kopf Einkommen

Anzahl	51
Arithm. Mittelwert	20934,5
Median	20625,0
Standardabweichungen	3018,79
Variationskoeffizient	14,4202%
Minimum	15853,0
Maximum	28766,0
Spannweite	12913,0
Stand. Schiefe	1,93465
Stand. Wölbung	0,421446

Der StatAdvisor
Diese Tabelle zeigt Summenstatistiken für Pro Kopf Einkommen. Sie enthält Maßzahlen für die zentrale Lage, die Variabilität und die Gestalt der Verteilung. Von speziellem Interesse sind hier die standardisierte Schiefe und die standardisierte Wölbung, die man verwenden kann, um herauszufinden, ob die Daten normalverteilt sind. Falls die Werte dieser Statistiken außerhalb des Bereiches von –2 bis +2 liegen, bedeutet das eine signifikante Abweichung von der Normalverteilung, wodurch ein statistischer Test (bei dem Normalverteilung unterstellt wird) (z.B.) mit Bezug zur Standardabweichung problematisch ist. In diesem Fall liegt der Wert für die standardisierte Schiefe innerhalb des Bereiches, den man für normalverteilte Daten erwarten würde. Der Wert für die standardisierte Wölbung liegt innerhalb des Bereiches, den man für normalverteilte Daten erwarten würde.

Abbildung 1.25: Maximiertes Ergebnisfenster mit der Summenstatistik-Tabelle

Die Tabelle zeigt einige interessante Statistiken auf. In den n = 51 Staaten (plus D.C.) reicht das Pro-Kopf-Einkommen von $15.853 bis $28.766. Das durchschnittliche Pro-Kopf-Einkommen beträgt $20.934,5 (arithmetrischer Mittelwert).

Unterhalb der Tabelle befindet sich das Ergebnis des *StatAdvisors*. Im vorliegenden Fall konzentriert sich dieses auf die zwei markierten Statistiken. Sie geben die Schiefe und die Wölbung der Daten an. Wie vom StatAdvisor erläutert, sollten die Schiefe- und Wölbung-Werte von Daten, die aus einer Normal- (oder Gauß-)Verteilung stammen, zwischen -2 und +2 liegen. In diesem Fall befinden sich beide Werte innerhalb dieses Bereichs. Dies deutet darauf hin, dass eine glockenförmige Kurve (Normalverteilung) das passende Modell für die Beobachtungen ist, auch wenn die Schiefe nahe an der statistischen Signifikanz ist.

Über einen Doppelklick in das Fenster gelangen Sie zurück in die vorhergegangene Anzeige. Maximieren Sie nun den Box-Whisker-Plot per Doppelklick in das entsprechende Ergebnisfenster:

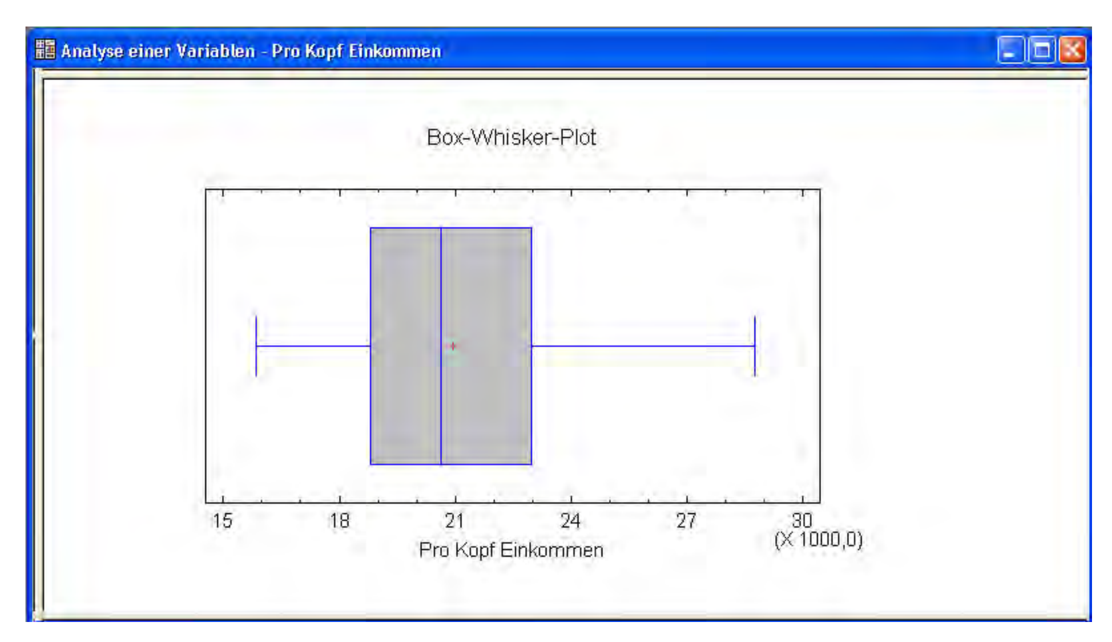


Abbildung 1.26: Maximiertes Ergebnisfenster mit dem Box-Whisker-Plot

Der Box-Whisker-Plot, der von John Tukey entwickelt wurde, gibt die Zusammenfassung der Daten einer Stichprobe mithilfe von 5 Zahlen wieder. Die Box überdeckt die mittleren 50 % der Daten, die sich vom unteren bis zum oberen Quartil erstrecken. Die horizontalen Linien (Whiskers) unterhalb und oberhalb der Box geben an, wo die kleinsten und die größten Werte der Stichprobe liegen. Die vertikale Linie innerhalb der Box bezeichnet den Median der Stichprobe, während das Pluszeichen (+) die Lage des Stichproben-Mittelwertes anzeigt. Dass der obere Whisker länger als der untere ist, während der Mittelwert größer als der Median ist, spiegelt die Rechts-Schiefe der Daten wieder.

1.6 Verwenden der Analyse-Symbolleiste

Wenn ein Analyse-Fenster wie das der *Analyse einer Variablen* erstmalig angezeigt wird, enthält es nur einige der verfügbaren Tabellen und Diagramme. Um weitere anzuzeigen, wählen Sie diese über die entsprechenden Schaltflächen der *Analyse-Symbolleiste*, die sich über dem Titel der Analyse befindet, aus:



Abbildung 1.27: Die Analyse-Symbolleiste

Die Schaltflächen auf der Analyse-Symbolleiste sind sehr wichtig. Die Aktionen, welche die sechs Symbole im linken Bereich verbildlichen, werden im Folgenden zusammengefasst:

	Name	*Funktion*
	Dateneingabe	Zeigt die Dateneingabe-Dialogbox an, so dass die ausgewählten Variablen geändert werden können.
	Analyse-Optionen	Zeigt Optionen für die Tabellen und Diagramme der vorliegenden Analyse an.
	Tabellen und Grafiken	Zeigt eine Liste mit den verfügbaren Tabellen und Diagrammen an.
	Ergebnisfenster-Optionen	Zeigt Optionen für die gerade maximierte Tabelle oder das gerade maximierte Diagramm an.
	Ergebnisse speichern	Speichert die errechneten Statistiken in Spalten eines Datenblatts.
	Diagramm-Optionen	Zeigt Optionen zur Bearbeitung eines maximierten Diagramms an wie Änderung des Titels, der Skalierung usw.

Abbildung 1.28: Wichtige Schaltflächen auf der Analyse-Symbolleiste

Weitere Schaltflächen im rechten Bereich stehen für andere Aktionen an einem maximierten Diagramm – sie werden in Kapitel 5 erläutert.

Wird beispielsweise die *Tabellen und Grafiken*-Schaltfläche gedrückt, so erscheint eine Dialogbox, die alle für die *Analyse einer Variablen* verfügbaren Tabellen und Diagramme anzeigt:

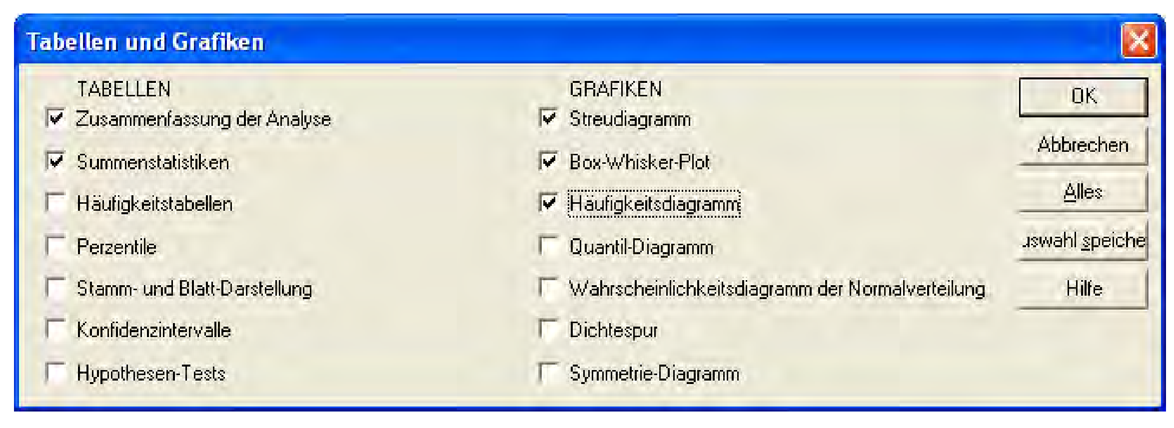


Abbildung 1.29: Liste der verfügbaren Tabellen und Diagramme

Hier markieren Sie z. B. das Kontrollkästchen *Häufigkeitsdiagramm* und verlassen die Dialogbox über *OK*. In der rechten Hälfte des Analyse-Fensters erscheint nun ein drittes Ergebnisfenster mit dem neuen Diagramm:

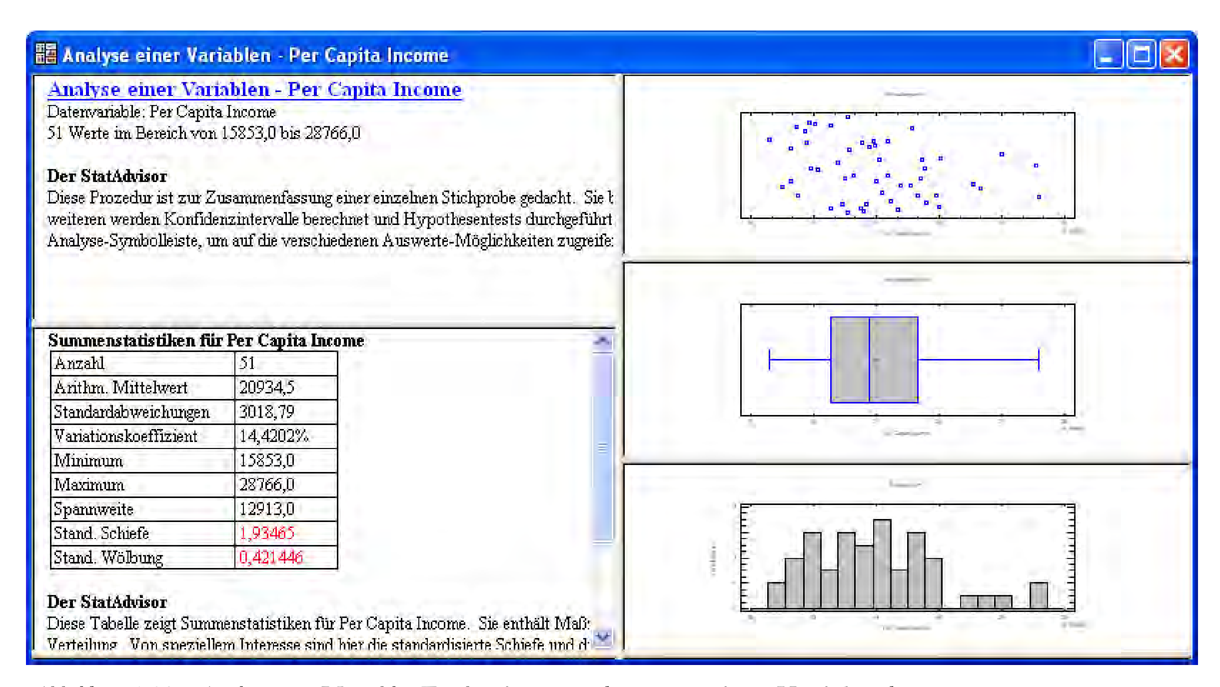


Abbildung 1.30: Analyse einer Variablen-Ergebnisfenster mit dem neu eingefügten Häufigkeitsdiagramm

Maximieren Sie nun das Häufigkeitsdiagramm, inderm Sie doppelt darauf klicken, und wählen Sie *Ergebnisfenster-Optionen* aus der Analyse-Symbolleiste. Es erscheint eine Dialogbox, welche die verfügbaren Optionen für das Diagramm anzeigt:



Abbildung 1.31: Ergebnisfenster-Optionen für das Häufigkeitsdiagramm

In der Box können die Anzahl der Säulen des Diagramms sowie der Bereich, über den sie reichen, verändert werden. Wenn Sie *Anzahl der Klassen* auf 15 setzen und *OK* wählen, wird das Diagramm dem entsprechend geändert:

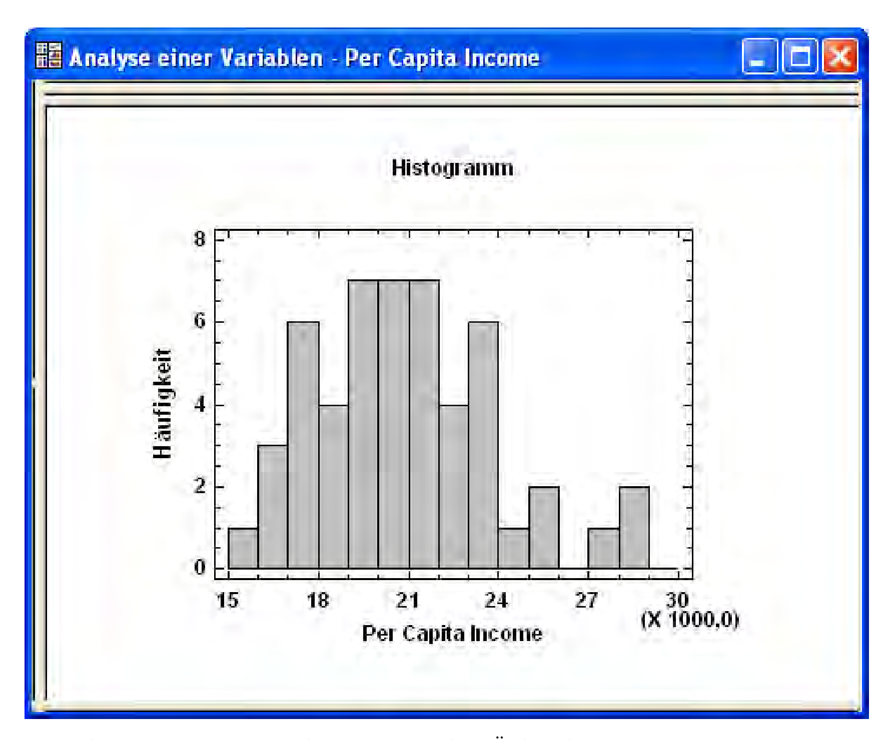


Abbildung 1.32: Häufigkeitsdiagramm nach dem Ändern der Klassenanzahl

Zusätzlich können Sie die Fülleffekte und/oder die Füllfarben der Säulen des Diagramms ändern, indem Sie *Diagramm-Optionen* wählen. In der angezeigten Dialogbox können Sie die meisten Eigenschaften des Diagramms einstellen. Wenn Sie hier das Register *Füllmuster* auswählen, wird Ihnen Folgendes angezeigt:

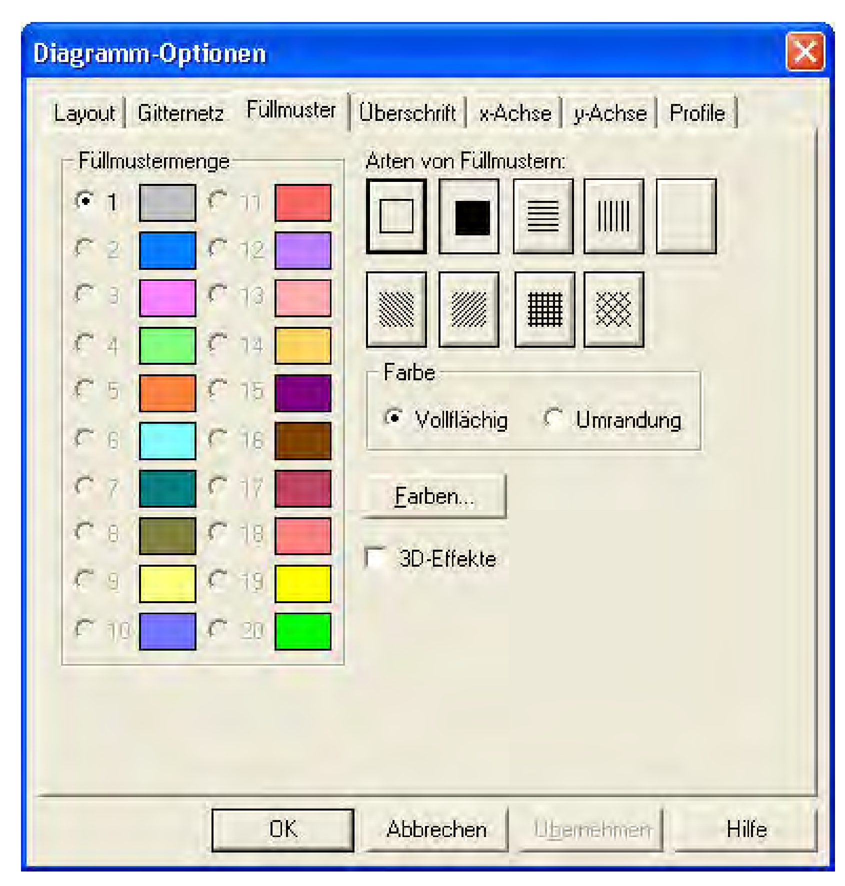


Abbildung 1.33: Dialogbox mit den Diagramm-Optionen

Sie können beispielsweise die Füllfarbe *1* und dann eine neue *Füllmusterart* oder *Farbe* auswählen, um das Erscheinungsbild der Säulen des Diagramms nach Ihren Wünschen zu modifizieren.

> ANMERKUNG: Die Aktionen der meisten Schaltflächen auf der Analyse-Symbolleiste können auch über die Kontextmenüs der Ergebnisfenster aufgerufen werden (rechter Mausklick in das Ergebnisfenster). In den Kontextmenüs sind alle verfügbaren Aktionen wählbar.

1.7 Verarbeiten der Ergebnisse

Nach der Auswertung der Daten können diese auf verschiedene Weisen weiter aufbereitet werden:

Aktion	*Methode*
Drucken der Ergebnisse.	Wählen Sie die *Drucken*-Schaltfläche auf der Symbolleiste, um alle Tabellen und Diagramme zu drucken, oder rufen Sie das Kontextmenü eines Ergebnisfensters auf (rechter Mausklick) und wählen Sie *Drucken*, um eine bestimmte Tabelle oder Diagramm zu drucken.
Darstellen der Ergebnisse in einem Webbrowser.	Wählen Sie *StatPublish* aus dem *Datei*-Menü. In der geöffneten Dialogbox können Sie eingeben, an welchen Ort das HTML-Dokument exportiert werden soll.
Einfügen der Ergebnisse in eine andere Anwendung.	Markieren Sie die gewünschte Grafik und wählen Sie *Kopieren* aus dem *Bearbeiten*-Menü. Öffnen Sie die andere Anwendung und fügen Sie die Grafik über *Bearbeiten – Einfügen* ein.
Sichern der Ergebnisse in einem Bericht.	Wählen Sie über das Kontextmenü der Grafik *Analyse in StatReporter kopieren*. Der StatReporter kann als RTF-Datei gesichert werden und in Programme wie Microsoft Word importiert werden (siehe Kapitel 7).
Sichern eines Diagramms als Bilddatei.	Maximieren Sie das zu sichernde Diagramm und wählen Sie *Grafik speichern* im *Datei*-Menü.

Abbildung 1.34: Methoden zur Weiterverarbeitung der Analyseergebnisse

Jede dieser Aktionen wird in einem der folgenden Kapitel ausführlich beschrieben.

1.8 Speichern der Arbeit

Sie können die laufende STATGRAPHICS Centurion XVI-Sitzung jederzeit sichern, indem Sie *Speichern – StatFolio speichern* aus dem *Datei*-Menü auswählen und einen Dateinamen eingeben:

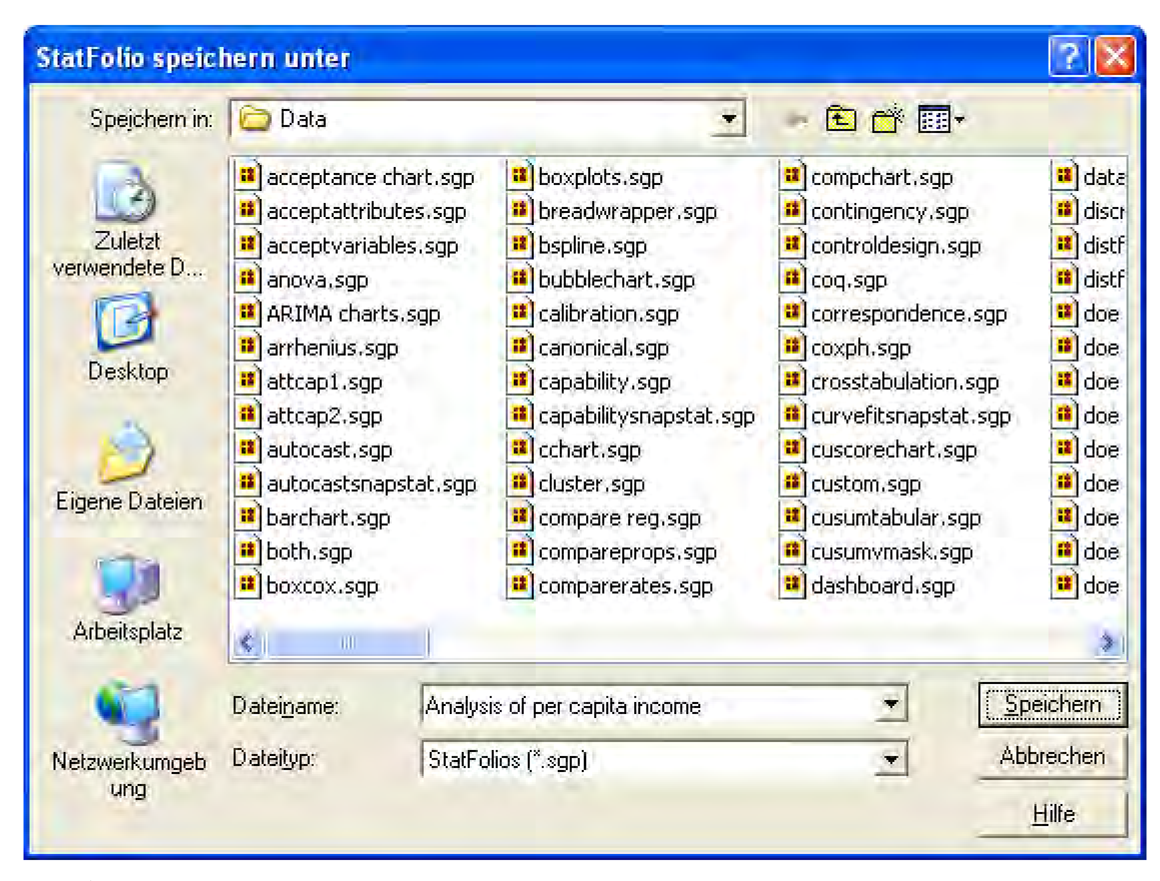


Abbildung 1.35: Dialogbox zum Speichern eines StatFolios

Ein StatFolio beeinhaltet alle Informationen, die benötigt werden, um die Analysen Ihrer aktuellen Sitzung wiederherzustellen, wie z. B. Verweise auf die Datenquellen, die Ihre Daten enthalten. Wenn Sie das StatFolio zu einem späteren Zeitpunkt erneut aufrufen, wird es die verwendeten Daten automatisch wieder einlesen und die Auswertungen erstellen. Jede Ihrer gewählten Optionen für die Analysen bleibt erhalten.

ANMERKUNG 1: Wenn sich die Datenquellen in der Zeit zwischen dem Sichern und dem erneuten Laden des StatFolios verändern, werden die entsprechenden

Auswertungen dementsprechend aktualisiert. So können Analysen, die regelmäßig benötigt werden, wiederholt ablaufen, ohne dass man sie erneut anlegen muss.

ANMERKUNG 2: Die Daten und das StatFolio werden in verschiedenen Dateien gespeichert. Wenn Sie ein StatFolio auf einen anderen Computer bringen, so sollten Sie unbedingt auch alle zugehörigen Datendateien mitnehmen.

Kapitel 2

Daten-Management

Zugreifen auf Daten aus Datendateien und Datenbanken, Transformieren von Datenwerten und Generieren strukturierter Daten.

Um in STATGRAPHICS Centurion XVI Daten auswerten zu können, müssen diese zunächst in das *Datenbuch* eingefügt werden. Das Datenbuch-Fenster enthält 26 Register mit *Daten-blättern*. Ein Datenblatt ist eine Matrix aus Zeilen und Spalten. Jede Spalte eines Datenblatts steht für ein messbares Merkmal (Variable), jede Zeile repräsentiert eine Beobachtung oder Messwert. Das folgende Datenblatt gibt Informationen zu verschiedenen Automarken und -modellen wieder:

C:\Programme\Statgraphics\STATGRAPHICS Centurion XVI\Data\93cars.sgd

	Make	Model	Type	Min Price	Mid Price
				price for basic version in $1,000	average of min and max prices in $1,000
1	Acura	Integra	Small	12.9	15.9
2	Acura	Legend	Midsize	29.2	33.9
3	Audi	90	Compact	25.9	29.1
4	Audi	100	Midsize	30.8	37.7
5	BMW	535i	Midsize	23.7	30
6	Buick	Century	Midsize	14.2	15.7
7	Buick	LeSabre	Large	19.9	20.8
8	Buick	Roadmaster	Large	22.6	23.7
9	Buick	Riviera	Midsize	26.3	26.3

93cars / testdatei2 / C / D / E / F / G / H / I / J / K

Abbildung 2.1: Beispiel-Datenblatt mit verschiedenen Automarken und -modellen

In diesem Kapitel wird beschrieben, was Sie über Daten in STATGRAPHICS Centurion XVI wissen müssen, z. B. wie sie abgerufen, bearbeitet und für statistische Analysen genutzt werden.

2.1 Das Datenbuch

Im STATGRAPHICS Centurion XVI-Datenblatt repräsentiert jede Spalte eine bestimmte Variable. Variablen sind gewöhnlich Merkmale oder (Mess-) Größen, denen die Eingaben aus den Zeilen des Datenblatts zugeordnet werden. Im *93cars*-Datenblatt gibt es z. B. eine Spalte, welche die Marken der Automobile identifiziert, sowie eine Spalte, die den Autotyp angibt, und Spalten, die den gemessenen Kraftstoffverbrauch (Miles per Gallon, MPG) der Fahrzeuge in der Stadt und auf der Autobahn enthalten. Des weiteren führen die Spalten die Länge, Höhe und das Gewicht der Fahrzeuge und ähnliche Informationen auf. Jede Spalte hat einen *Namen* sowie einen vorher zugewiesenen *Typ*. Der *Name* dient zum Aufrufen der Daten für die Analysen, während der *Typ* aussagt, wie die Daten analysiert werden. Zudem kann eine Spalte einen *Kommentar* enthalten, der zusätzliche Information über ihren Inhalt bereitstellt. HINWEIS: Die obigen Daten wurden vom „Journal of Statistical Education Data Archive" bezogen (www.amstat.org/publications/jse/jse_data_archive.html) und werden mit dessen Erlaubnis verwendet.

Um die Eigenschaften einer Spalte anzuzeigen oder zu ändern, können Sie über einen Doppelklick in den Kopfbereich der Spalte die Dialogbox *Spalten bearbeiten* öffnen:



Abbildung 2.2: Dialogbox zum Eingeben der Spalteneigenschaften

Sie können folgende Eigenschaften festlegen:

Name: mit 1 bis 32 Zeichen. Über den Namen identifizieren Sie die Spalten, die in eine statistische Auswertung einbezogen werden sollen. Innerhalb eines Datenblatts muss jede Spalte einen individuellen Namen haben, jedoch können Spalten in verschiedenen Datenblättern die gleichen Namen besitzen. Namen können alle Zeichen, auch Leerzeichen, enthalten. Variablenamen unterscheiden **keine** Groß- und Kleinschreibung.

1. **Kommentar:** mit 0 bis 64 Zeichen. Enthält weitere Information zum Inhalt der Spalte.

2. **Typ:** legt den in der Spalte zugelassenen Datentyp fest. Folgende Typen können bestimmt werden:

Typ	*Inhalte*	*Beispiel*
Numerisch	Jede gültige Zahl	3,14
Text	Eine alphanumerische Zeichenfolge	Chevrolet
Integer	Eine ganze Zahl	105
Datum	Tag, Monat und Jahr	30.04.2005
Monat	Monat und Jahr	04.2005
Quartal	Quartal und Jahr	Q2.2005
Zeit (HH:MM)	Stunde und Minute	3:15
Zeit (HH:MM:SS)	Stunde, Minute und Sekunde	3:15:53
Datum-Zeit (HH:MM)	Tag, Monat, Jahr, Stunde und Minute	30.04.2005 3:15
Datum-Zeit (HH:MM:SS)	Tag, Monat, Jahr, Stunde, Minute und Sekunde	30.04.2005 3:15:53
Festkommazahl	Zahl mit fester Anzahl an Nachkommastellen (1-9)	34,10
Formel	Aus anderen Spalten berechnet	MPG City/MPG Highway

Abbildung 2.3: Verfügbare Variablentypen

Wenn Sie Daten in ein Datenblatt eingeben, müssen diese mit dem angegebenen Typ der Spalte übereinstimmen. Wenn Sie z. B. versuchen, Buchstaben in eine numerische Spalte einzugeben, wird dies abgelehnt werden. Bei der Dateneingabe muss zudem das Format der Daten den aktuellen Windows-Einstellungen entsprechen. STATGRAPHICS Centurion XVI akzeptiert die aktuellen Windows-Einstellungen für:

1. Dezimalenzeichen für numerische Werte
2. Zeitformat and Trennzeichen für Zeitangaben

3. Datum-Kurzformat und Trennzeichen für Datum

Die Einstellungen Ihres Systems können Sie in der Windows-*Systemsteuerung* überprüfen.

Bei Eingabe eines Datums müssen Sie das Format verwenden, das in der Dialogbox *Bearbeiten – Voreinstellungen* bestimmt ist: 4-stellige Jahreszahlen (wie bei 30.04.2005) oder 2-stellige Jahreszahlen (wie bei 30.04.05). Bei einer 2-stelligen Jahreszahl wird angenommen, dass diese zwischen 1950 und 2049 liegt. Informationen zum Typ *Formel* finden Sie weiter unten in diesem Kapitel unter *Bearbeiten der Daten*.

2.2 Zugreifen auf die Daten

In Kapitel 1 haben wir gezeigt, wie Daten manuell in ein Datenblatt eingegeben werden. Häufig greifen Anwender jedoch auf Daten zurück, die bereits in einer anderen Datei oder Anwendung vorliegen. Es gibt drei einfache Wege, solche bereits existierenden Daten in ein STATGRAPHICS Centurion XVI-Datenblatt einzufügen:

1. **Lesen einer bestehenden Datendatei:** Sie können bereits in einer Datei erfasste Daten in das Datenblatt einlesen, indem Sie *Datei – Öffnen – Datenquelle öffnen* wählen. So können verschiedene Dateiformate wie z. B. Excel-Dateien, ASCII-Dateien mit Trennzeichen, XML-Dateien, STATGRAPHICS-Dateien und Dateien aus anderen Statistikprogrammen gelesen werden.

2. **Kopieren und Einfügen mit der Windows-Zwischenablage:** Daten aus einem Programm wie Excel können über die Windows-Zwischenablage kopiert und in STATGRAPHICS Centurion XVI über *Bearbeiten – Einfügen* eingesetzt werden.

3. **Abfragen von SQL-Datenbanken:** Befinden sich die Daten in einer ODBC-kompatiblen Datenbank wie Oracle oder Microsoft Access, können sie über *Datei – Öffnen – Datenquelle öffnen* und *ODBC-Abfrage* aufgerufen werden.

2.2.1 Lesen der Daten aus einer STATGRAPHICS Centurion-Datendatei

Um Daten zu nutzen, die bereits in einer STATGRAPHICS Centurion XVI-Datendatei gespeichert sind, markieren Sie eins der 26 Datenblätter des Datenbuchs, indem Sie auf das entsprechende Register klicken. Wählen Sie dann *Datei – Öffnen – Datenquelle öffnen* und markieren Sie das Optionsfeld *STATGRAPHICS-Datendatei* in der Dialogbox:

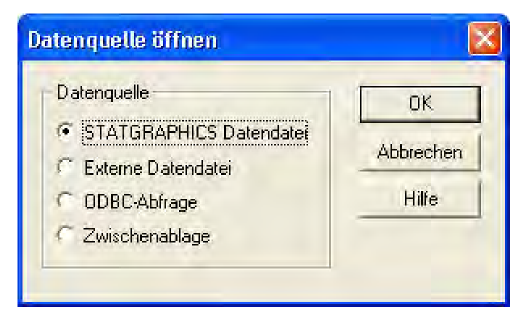


Abbildung 2.4: Dialogbox zum Öffnen einer Datenquelle

Wählen Sie *OK* und markieren Sie im nächsten Dialog die gewünschte STATGRAPHICS-Datei:

Abbildung 2.5: Wählen einer STATGRAPHICS-Datendatei

Es können Datendateien von STATGRAPHICS Centurion XVI und jeder Vorgängerversion von STATGRAPHICS, einschließlich STATGRAPHICS *Plus*, gelesen werden. Die Daten aus der Datei ersetzen alle Inhalte des vorher markierten Datenblatts.

2.2.2 Lesen der Daten aus Excel-, ASCII-, XML- oder anderen Datendateien

Um Daten aus externen Datendateien zu lesen, markieren Sie eins der 26 Datenblätter des Datenbuchs, indem Sie das entsprechende Register anklicken. Über *Datei – Öffnen – Datenquelle öffnen* rufen Sie die folgende Dialogbox auf und wählen *Externe Datendatei*:

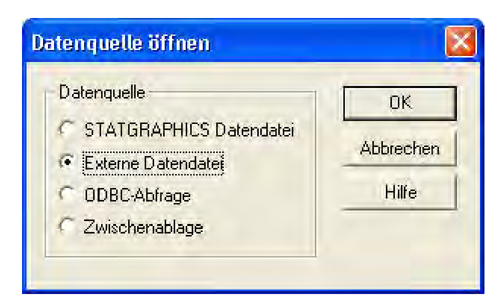


Abbildung 2.6: Dialogbox zum Öffnen einer Datenquelle

Nachdem Sie *OK* angeklickt habe wird eine Dialogbos angezeigt, über die Sie die zu importierende Datei und weitere relevante Informationen angeben können:

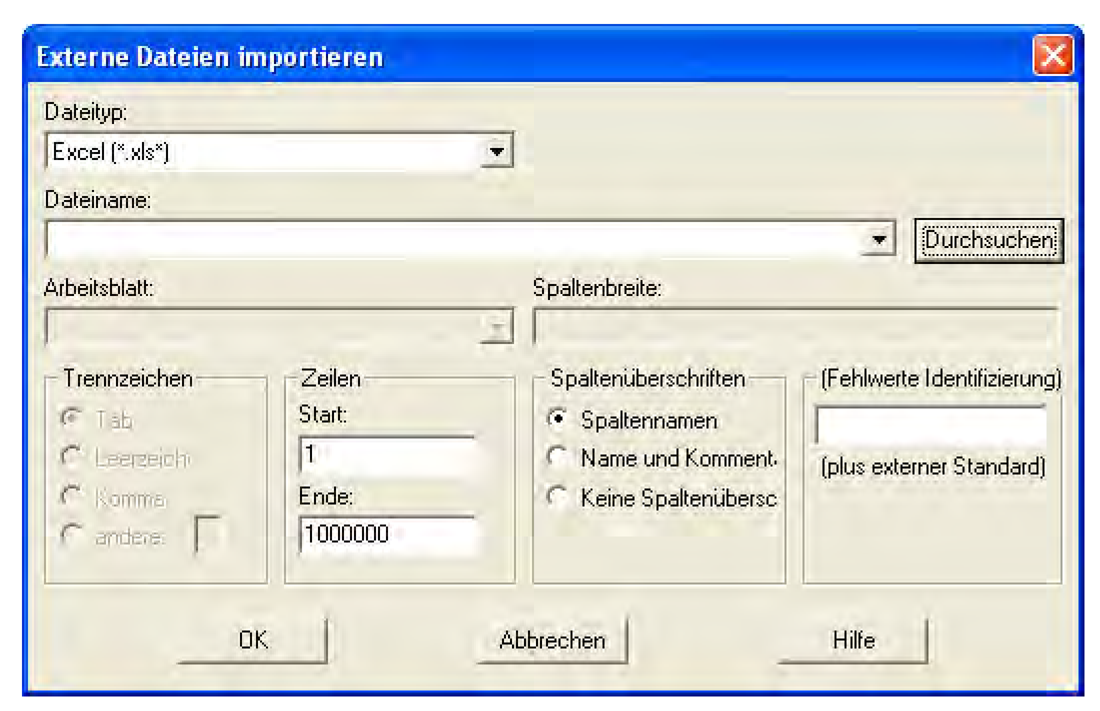


Abbildung 2.7: Wählen einer externen Datendatei

Die Dialogbox enthält folgende Felder:

1. **Dateityp** – Typ der zu importierenden Datei. STATGRAPHICS Centurion XVI kann Daten aus zahlreichen Programmen importieren, einschließlich Excel, Matlab, Minitab, JMP, SPSS, SAS, und viele andere Statistikprogramme.

2. **Dateiname** – Name der zu importierenden Datei. Klicken Sie die Schaltfläche *Durchsuchen* an, um die gewünschte Datei auszuwählen.

3. **Arbeitsblatt** – Name des zu importierenden Arbeitsblattes (falls vorhanden).

4. **Spaltenbreite** – Breite jeder Spalte, getrennt duch Kommas (nur bei formatierten ACII-Dateien).

5. **Trennzeichen** – Trennzeichen zwischen den Spalten (nur bei formatierten ASCII-Dateien).

6. **Zeilen** – Bereich der Zeilen des Arbeitsblattes, der eingelesen werden soll. Dieser Bereich schließt evtl. vorhandene Variablennamen und Kommentare mit ein.

7. **Spaltenüberschriften** – Informationen, die in den ersten zwei Zeilen des angegebenen Bereichs enthalten sind (bei Spreadsheet-Programmen wie z. B. Excel). Diese zwei direkt über den einzulesenden Daten befindlichen Zeilen können Spaltennamen und/oder Kommentare enthalten. Falls keine Spaltennamen enthalten sind, werden Standardnamen erzeugt.

8. **Fehlwerte-Identifizierung** – alle speziellen Zeichen, die in der externen Datei als Platzhalter für fehlende Daten verwendet werden, z. B. *NA*. Zellen, die den hier angegebenen Wert enthalten, wandelt STATGRAPHICS Centurion XVI in leere Zellen um.

Wählen Sie *OK*, um die Daten aus der externen Datei in STATGRAPHICS Centurion XVI einzulesen. Dabei wird jede Spalte einzeln geprüft und ihr anschließend ein entsprechender Spaltentyp zugewiesen. Die Daten können nun analysiert werden.

2.2.3 Übertragen von Daten durch *Kopieren* und *Einfügen*

Die einfachste Möglichkeit, Daten aus einer Anwendung nach STATGRAPHICS Centurion XVI zu transportieren, bietet häufig die Windows-Zwischenablage. Dies betrifft z. B. Daten aus Excel-Dateien. Um Excel-Daten zu übertragen, öffnen Sie das Excel-Programm und die gewünschte Datendatei. Markieren Sie die benötigten Daten und kopieren Sie diese über *Bearbeiten – Kopieren* in die Zwischenablage. In STATGRAPHICS Centurion XVI setzen Sie die Daten dann über *Bearbeiten – Einfügen* direkt in das Datenblatt ein. Während des Einfügens von

Daten in eine Spalte des Datenblatts überprüft STATGRAPHICS Centurion XVI die Daten automatisch und weist der Spalte den passenden Spaltentyp zu.

Beim Kopieren und Einfügen von Daten können auch die Spaltennamen sowie die zugehörigen Kommentare übertragen werden. Beziehen Sie auch die Spaltennamen und Kommentare in Excel ein, wenn Sie Daten in die Zwischenablage kopieren. Im STATGRAPHICS Centurion XVI-Datenblatt markieren Sie dann, bevor Sie *Einfügen* wählen, die Kopfzeile der Matrix. Es werden die Informationen, die in der Zwischenablage an erster Stelle stehen, in die Kopfzeile(n) des Datenblatts übertragen.

2.2.4 Abfragen einer ODBC-Datenbank

Mit STATGRAPHICS Centurion XVI können auch Daten aus einer Oracle-, Access- oder anderen ODBC-fähigen Datenbank gelesen werden. Um auf die Daten einer Datenbank zuzugreifen, wählen Sie zunächst *Datei – Öffnen – Datenquelle öffnen* und anschließend *ODBC-Abfrage* in der Dialogbox:

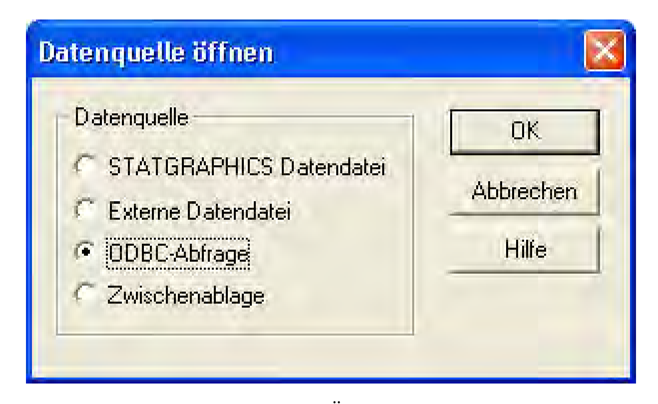


Abbildung 2.8: Dialogbox zum Öffnen einer Datenquelle

Es folgt nun eine Reihe von Dialogboxen, mit denen Sie:

1. den Namen der zu lesenden Datenquelle wählen.
2. die zu übertragenden Felder wählen.
3. einen Filter festlegen, um die Zahl der abzufragenden Datensätze einzugrenzen.
4. eine Art der Anordnung der Ergebnisse bestimmen.

Hieraus wird eine SQL-Abfrage erstellt, und die Ergebnisse werden in das aktive STATGRAPHICS Centurion XVI-Datenblatt eingesetzt. Weitere Informationen zur Gestaltung von ODBC-Abfragen entnehmen Sie bitte dem PDF-Dokument *Data Files and StatLink*.

2.3 Bearbeiten der Daten

Sobald die Daten in das STATGRAPHICS Centurion XVI-Datenblatt eingefügt wurden, können sie auf verschiedene Arten bearbeitet werden:

1. Die Daten können kopiert und an einer anderen Stelle eingefügt werden.

2. Aus den bestehenden Spalten können weitere Spalten erstellt werden.

3. Die Daten können mittels eines algebraischen Ausdrucks oder einer mathematischen Funktion transformiert werden.

4. Das Datenblatt kann nach einer oder mehreren Spalten sortiert werden.

5. Datenwerte können rekodiert werden, z. B. um Gruppen zu bilden.

6. Daten, die sich in mehreren Spalten befinden, können in einer einzigen Spalte neu angeordnet werden, wenn dies für eine statistische Prozedur notwendig ist.

Die Umsetzung dieser wichtigen Vorgänge wird in den folgenden Abschnitten beschrieben.

2.3.1 *Kopieren* und *Einfügen* der Daten

Das STATGRAPHICS Centurion XVI-Datenblatt unterstützt viele der typischen Spreadsheet-Bearbeitungsoptionen, wie z. B. *Ausschneiden*, *Kopieren*, *Einfügen*, *Einsetzen* und *Entfernen*. Bei der Anwendung dieser Optionen ist es wichtig, zu bedenken, dass jede Spalte einen bestimmten Spaltentyp besitzt: Wenn Sie versehentlich *Text*-Daten in eine numerische Spalte einfügen, passt STATGRAPHICS Centurion den Spaltentyp automatisch an die nun vorhandenen Daten an. Um den Spaltentyp einer Spalte zu bestimmen, klicken Sie den Spaltenkopf der entsprechenden Spalte doppelt an, um die Dialogbox *Spalten bearbeiten* zu öffnen. Hier wird der Spaltentyp angezeigt und kann geändert werden.

2.3.2 Erstellen neuer Variablen aus bestehenden Spalten

STATGRAPHICS Centurion XVI verfügt über einen breiten Bereich an Operatoren, die Sie bei mathematischen Berechnungen unterstützen. Zu den wichtigsten Anwendungen der Operatoren bei der Analyse von Daten gehört die Erstellung neuer Variablen basierend auf bereits bestehenden Spalten. In STATGRAPHICS Centurion XVI werden neue Variablen wie folgt erstellt:

1. „On-the-fly", d. h. direkt in den Datenfeldern der Dateneingabe-Dialogboxen, ohne die Variablen im Datenblatt zu speichern.

2. Durch Erstellen einer neuen Spalte in einem der 26 Datenblätter des Datenbuches.

Angenommen, es soll für jedes der Fahrzeuge der *93cars*-Datendatei das Verhältnis der "Reichweite" (Miles Per Gallon, *MPG*) in der Stadt zu der auf der Autobahn betrachtet werden. Die Datei enthält die zwei separaten Spalten *MPG City* und *MPG Highway*. Um die Verteilung dieses Quotienten zu analysieren, können Sie z. B. die *Analyse einer Variablen* wählen. Den gewünschten Quotienten geben Sie direkt in das Eingabefeld *Daten* in der Dialogbox ein:



Abbildung 2.9: Erstellen einer On-The-Fly-Umformung

Wenn Sie *OK* wählen, wird eine Auswertung für die transformierten Daten erzeugt, ohne dass die Daten im Datenblatt deswegen geändert werden:

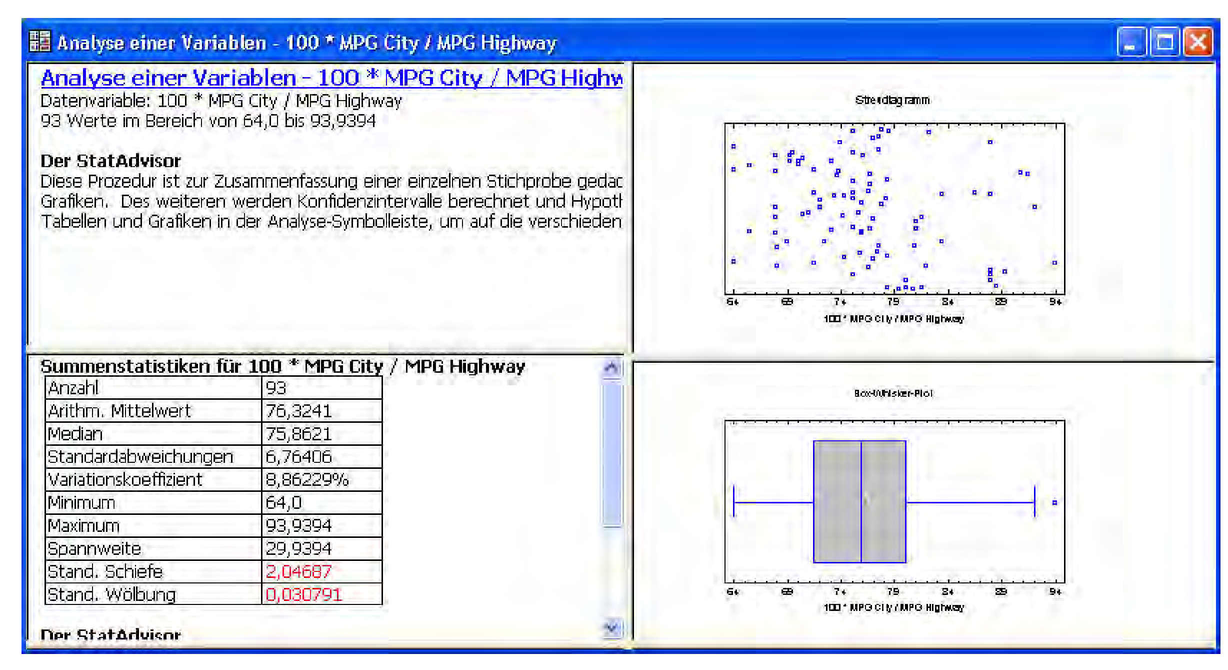


Abbildung 2.10: Analyse einer Variablen mit den transformierten Daten

Das durchschnittliche Verhältnis beträgt 76,3 %, wobei der kleinste Wert bei 64,0 % und der höchste bei 93,9 % liegt. Die Fähigkeit von STATGRAPHICS, viele Auswertungen ohne eine Veränderung der Daten im Datenblatt vorzunehmen, trägt dazu bei, die Zusammenhänge in Ihren Daten schnell und unkompliziert zu erklären.

Wenn gewünscht, kann auch eine neue Spalte mit den ermittelten Werten im Datenblatt angelegt werden. Hierzu können Sie in dem Fenster mit der geöffneten *93cars*-Datei die Spalte *Col_27* über einen Doppelklick in den Spaltenkopf auswählen und die Dialogbox *Spalten bearbeiten* aufrufen. In der Box geben Sie eine Variable des Typs *Formel* mit der oben angegebenen Umformung an:

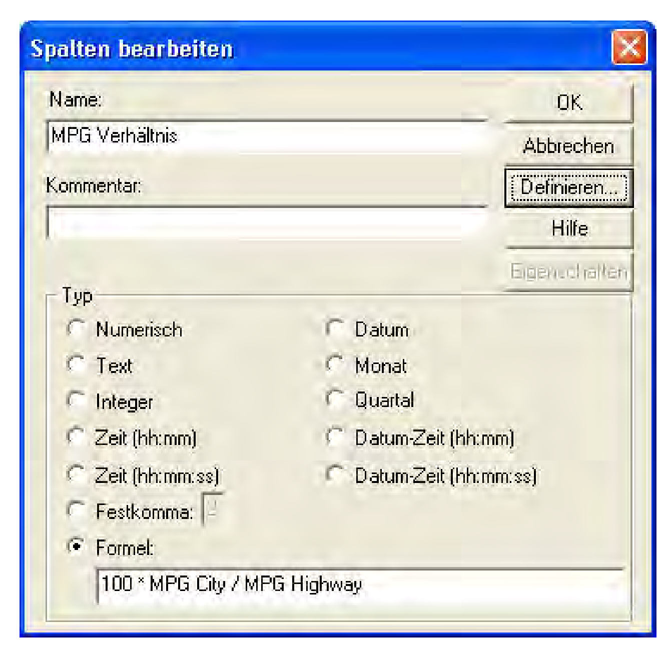


Abbildung 2.11: Erstellen einer Formel-Spalte

Sie erstellen so eine neue Spalte mit dem MPG-Verhältnis, deren Werte aus den bereits bestehenden Spalten *MPG City* und *MPG Highway* berechnet werden. *Formel*-Spalten enthalten im Datenblatt graue Zeichen, da ihre Werte automatisch aus anderen Spalten berechnet werden:

C:\Programme\Statgraphics\STATGRAPHICS Centurion XVI\Data\93cars.sgd

	Luggage	Weight	Domestic	MPG Ratio
	cu. ft.	pounds	1=U.S. manufacturer	
1	11	2705	0	80,6451612903
2	15	3560	0	72
3	14	3375	0	76,9230769231
4	17	3405	0	73,0769230769
5	13	3640	0	73,3333333333
6	16	2880	1	70,9677419355
7	17	3470	1	67,8571428571
8	21	4105	1	64
9	14	3495	1	70,3703703704
10	18	3620	1	64
11	14	3935	1	64

93cars / B / C / D / E / F / G / H / I / J / K

Abbildung 2.12: Erscheinungsbild einer Formel-Spalte in einem Datenblatt

Sobald sich die Werte in den Spalten *MPG City* oder *MPG Highway* ändern, werden die Werte in der Spalte *MPG Verhältnis* automatisch aktualisiert.

ANMERKUNG: Die Neuberechnung der Formel-Spalten geschieht erst, wenn die Daten dieser Spalten für weitere Berechnungen benötigt, gespeichert oder gedruckt werden. Sie können die Neuberechnung direkt anfordern, indem Sie im *Bearbeiten*-Menü *Formeln aktualisieren* auswählen.

2.3.3 Umformen der Daten

In STATGRAPHICS Centurion XVI ist eine große Anzahl mathematischer Funktionen enthalten, die zum Umformen bestehender Daten angewendet werden können. In der Weise, wie Sie eine neue Variable erstellen, so können Sie auch Tranformationen direkt über die Eingabe-Dialogbox eines Datenfelds oder durch das Erstellen einer neuen Spalte im Datenblatt vornehmen.

Angenommen, die Reichweite eines Fahrzeugs soll über dem natürlichen Logarithmus des Fahrzeuggewichts aufgetragen werden. Die Auswahl von *x-y-Diagramm* aus dem Hauptmenü öffnet die folgende Dateneingabe-Dialogbox :



Abbildung 2.13: Datentransformation in einer Dateneingabe-Dialogbox

Anstelle des Spaltennamens können Sie auch einen STATGRAPHICS Centurion-Ausdruck in ein Datenfeld eingeben. STATGRAPHICS Centurion-Ausdrücke sind Formeln, die mit den Daten unter Verwendung von algebraischen Symbolen und speziellen Operatoren arbeiten. STATGRAPHICS Centurion XVI verwendet eine große Bandbreite an Operatoren, die in dem PDF-Dokument *STATGRAPHICS Operators* beschrieben werden. Folgende Tabelle listet die gebräuchlichsten Operatoren auf:

Operator	*Verwendung*	*Beispiel*
+	Addition	X+100
-	Subtraktion	X-100
/	Division	X/100
*	Multiplikation	X*100
^	Potenzierung	X^2
ABS	Absolutwert	ABS(X)
AVG	Durchschnitt	AVG(X)
DIFF	Differenzbildung	DIFF(X)
EXP	Exponentialfunktion	EXP(10)
LAG	Lag von k Perioden	LAG(X;k)
LOG	Natürlicher Logarithmus	LOG(X)
LOG10	Logarithmus zur Basis 10	LOG10(X)
MAX	Maximum	MAX(X)
MIN	Minimum	MIN(X)
SD	Standardabweichung	SD(X)
SQRT	Quadratwurzel	SQRT(X)
STANDARDIZE	Standardisieren (z-Werte)	STANDARDIZE(X)

Abbildung 2.14: Häufig verwendete STATGRAPHICS-Operatoren

In einem STATGRAPHICS Centurion-Ausdruck können die Operatoren – im Einklang mit den üblichen algebraischen Rechengesetzen – kombiniert werden. Der folgende Ausdruck rechnet z.B. jeden Wert der Spalte *Weight* in den angegebenen Bruch um:

(*Weight* – MIN(*Weight*)) / (MAX(*Weight*) - MIN(*Weight*))

Die Klammern stellen dabei sicher, dass die Subtraktionen vor der Division ausgeführt werden. Ausdrücke unterscheiden nicht zwischen Groß- und Kleinbuchstaben, Leerzeichen sind nicht relevant.

Jede Dateneingabe-Dialogbox enthält eine *Transformieren*-Schaltfläche wie in *Abbildung 2.13*, die Ihnen helfen kann, STATGRAPHICS Centurion-Ausdrücke zu formulieren, wenn Sie unsicher sind, welcher Operator in einem bestimmten Fall angewendet werden sollte. Setzen Sie den Cursor der Maus in ein Datenfeld in der Dialogbox und wählen Sie *Transformieren*, um eine Dialogbox, die der folgenden ähnelt, aufzurufen:



Abbildung 2.15: Dialogbox nach Anklicken der Schaltfläche Transformieren

In der rechten Hälfte befindet sich eine Liste aller STATGRAPHICS Centurion-Operatoren mit einem Hinweis auf die Anzahl der jeweils anzugebenden Argumente in Klammern. Wählen Sie einen Operator mit der linken Maustaste aus, um ihn das Eingabefeld *Ausdruck* einzufügen. Nachdem Sie die Fragezeichen (?) durch die entsprechenden Spaltennamen oder durch Zahlen ersetzt haben, können Sie *Anzeigen* wählen, um zunächst die ersten durch den Ausdruck generierten Werte anzuzeigen, oder Sie wählen *OK*, um den Ausdruck in die Dateneingabe-Dialogbox einzufügen.

ANMERKUNG: Sie können die Ausdrücke auch manuell in der Dateneingabe-Dialogbox eingeben. Die Verwendung der *Transformieren*-Box ist optional.

Wenn die Transformation in der Dateneingabe-Dialoxbox, wie z. B. in *Abbildung 2.13* ersichtlich, eingegeben ist, wird diese beim Ausführen der Prozedur verwendet:

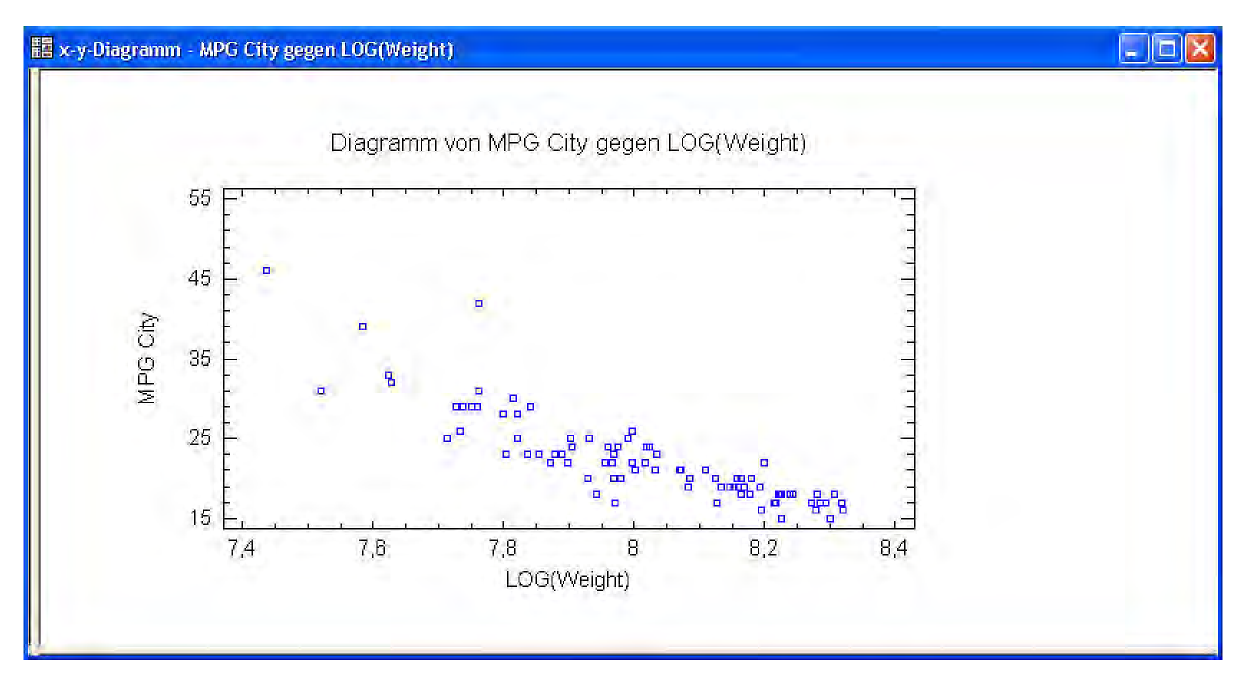


Abbildung 2.16: x-y-Diagramm unter Verwendung der transformierten Gewichtswerte

In ähnlicher Weise können STATGRAPHICS Centurion-Operatoren auch bei dem Erstellen von *Formel*-Spalten eingesetzt werden.

2.3.4 Sortieren der Daten

Die Inhalte eines Datenblatts können sortiert werden, indem Sie die Spalte(n) markieren, die für die Sortierreihenfolge verwendet werden sollen, und anschließend *Daten sortieren* im *Bearbeiten*-Menü auswählen. Um z.B. die Daten der *93cars*-Datei nach der Reichweite (Miles Per Gallon, MPG) zu sortieren, markieren Sie die Spalten *MPG City* sowie *MPG Highway* und wählen dann *Daten sortieren*. Damit öffnen Sie folgende Dialogbox:



Abbildung 2.17: Dialogbox mit den Optionen zum Sortieren

Sie können eine oder zwei Spalten festlegen, nach denen und in welcher Reihenfolge sortiert werden soll. Das Sortieren nach *MPG City* und anschließend nach *MPG Highway* sortiert zuerst nach dem Kraftstoffverbrauch *MPG City* und dann – bei Fahrzeugen mit gleichen Werten in *MPG City* – nach dem Kraftstoffverbrauch *MPG Highway*:

C:\Programme\Statgraphics\STATGRAPHICS Centurion XVI\Data\93cars.sgd

	Mid Price	Max Price	MPG City	MPG Highway	
	average of min and max prices in $1,000	price for a premium version in $1,000	miles per gallon in city driving	miles per gallon in highway driving	o:
1	16.6	18.6	15	20	0
2	19.9	25.3	15	20	1
3	23.7	24.9	16	25	1
4	34.7	36.3	16	25	1
5	40.1	42.7	16	25	2
6	19	24.4	17	21	1
7	19.7	22.7	17	21	0
8	47.9	50.4	17	22	1
9	19.1	21.5	17	23	0
10	38	41.5	17	25	1
11	32.5	32.5	17	25	1
12	18.8	19.6	17	26	1
13	34.3	35.3	17	26	2
14	22.7	26.6	18	22	1

93cars B C D E F G H I J K

Abbildung 2.18: 93cars.sgd-Datei nach dem Sortieren

ANMERKUNG: Für die Anwendung statistischer Prozeduren müssen die Daten im Vorfeld nicht sortiert werden, da dies, soweit notwendig, automatisch ausgeführt wird. Auch wird die Datendatei auf der Festplatte durch das Sortieren nicht verändert, bis Sie die Änderungen speichern. Das Sortieren betrifft also nur die Abfolge der Zeilen, mit der sie im Datenblatt angezeigt werden.

2.3.5 Rekodieren der Daten

Mitunter kann es hilfreich sein, Daten zu rekodieren, um sie in Gruppen anzuordnen oder um ihnen neue Beschriftungen zuzuweisen. Um Daten in einer Spalte zu rekodieren, klicken Sie in den Spaltenkopf der betroffenen Spalte und wählen *Daten rekodieren* im *Bearbeiten*-Menü. Folgende Dialogbox wird hiermit geöffnet :



Abbildung 2.19: Dialogbox zum Rekodieren der Daten

In der *93cars*-Datei enthält z.B. die Spalte *Domestic* eine *1* für jedes Fahrzeug, das von U.S.-amerikanschen Herstellern produziert wurde, allen anderen Autos ist eine *0* zugewiesen. Um alle *0*-Werte in der Spalte durch *Ausländisch* und alle *1*-Werte durch *USA*. zu ersetzen, kann die obige Dialogbox verwendet werden. Hier können Sie bis zu sieben Werte-Bereiche auf einmal rekodieren.

Weitere Beispiele und Erläuterungen zum Rekodieren enthält das PDF-Dokument *Edit Menu*.

2.3.6 Zusammenführen mehrerer Spalten

Viele der statistischen Prozeduren in STATGRAPHICS Centurion XVI setzen voraus, dass sich die zu analysierenden Daten in einer einzigen Spalte befinden. Dies ist jedoch nicht immer der Fall. Angenommen, eine einfache Stichprobe mit 12 Beobachtungen wurde wie folgt in vier Spalten erfasst:

<unbezeichnet>

	Col_1	Col_2	Col_3	Col_4	Col_5	Col_6	Col_7
1	5	2	8	9			
2	2	4	7	2			
3	7	2	3	1			
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							

A / B / C / D / E / F / G / H / I / J

Abbildung 2.20: Beispiel-Daten, verteilt auf mehrere Spalten

Sie könnten die Daten über viele *Kopieren*- und *Einfügen*-Vorgänge in einer einzigen Spalte zusammenführen. Eine einfachere Lösung wäre es jedoch, die Prozedur *Spalten zusammenfassen* auszuführen. *Spalten zusammenfassen* befindet sich im klassischen Menü unter *Beschreiben*. Um die Prozedur zu initiieren, geben Sie zunächst die Namen der zu analysierenden Spalten in die aufgerufene Dateneingabe-Dialogbox ein:

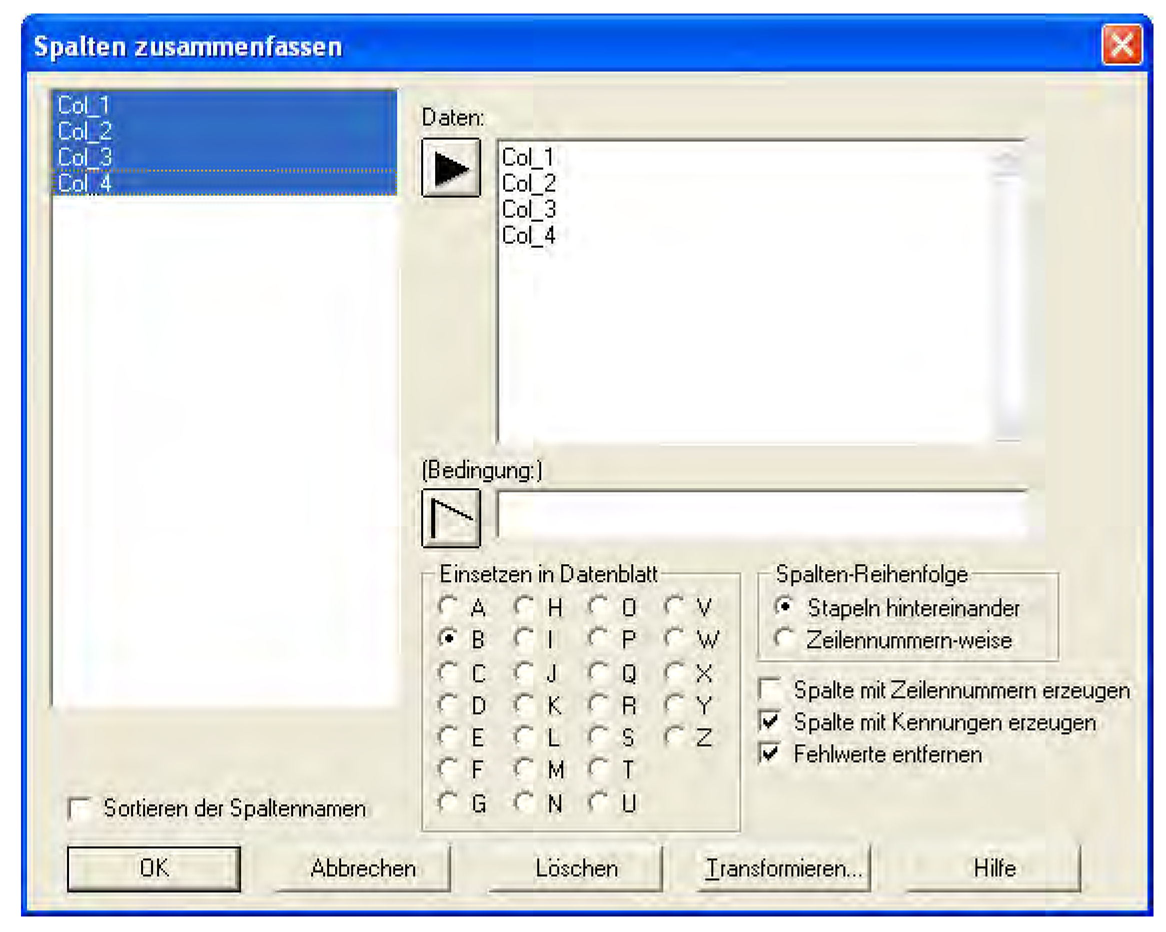


Abbildung 2.21: Dateneingabe-Dialogbox für ***Spalten zusammenfassen***

Die Dialogbox enthält folgende Felder:

1. **Daten** – zusammenzuführende Spalten.

2. **Bedingung** – Standardfeld zur Angabe einer Untergruppe von Zeilen.

3. **Einsetzen in Datenblatt** – Ziel-Datenblatt für die zusammengefassten Daten.

4. **Spalten-Reihenfolge** – gibt an, ob die Daten spalten- oder zeilenweise gestapelt werden sollen (eine Spalte/Zeile nach der anderen).

5. **Spalte mit Zeilennummern erzeugen** – gibt an, das seine weitere Spalte erstellt wird, die anzeigt, in welcher der ursprünglichen Zeilen die Werte enthalten waren.

6. **Spalte mit Kennungen erzeugen** – gibt an, das eine zweite Spalte erstellt wird, die anzeigt, in welcher ursprünglichen Spalten die Werte enthalten waren.

7. **Fehlwerte entfernen** – leere Zellen werden übersprungen anstatt einen Platzhalter einzufügen.

Nachdem Sie *OK* angeklickt haben, werden die Daten wie folgt in einer einzigen Spalte zusammengefasst:

<unbezeichnet>

	Zusammengeführt	Identifier	Col_3
1	5	Col_1	
2	2	Col_1	
3	7	Col_1	
4	2	Col_2	
5	4	Col_2	
6	2	Col_2	
7	8	Col_3	
8	7	Col_3	
9	3	Col_3	
10	9	Col_4	
11	2	Col_4	
12	1	Col_4	
13			
14			

A B C D E F

Abbildung 2.22: In einer einzigen Spalte zusammengefasste Daten

2.4 Generieren von Daten

Mit STATGRAPHICS Centurion XVI können Daten generiert und in die Spalten des Datenblatts eingesetzt werden. Dieser Abschnitt beschreibt zwei hierfür wichtige Möglichkeiten:

1. Generieren von Daten mit einfachen Strukturen.

2. Generieren von Zufallszahlen.

2.4.1 Generieren strukturierter Daten

Viele der Prozeduren in STATGRAPHICS Centurion XVI, insbesondere der für die Varianzanalyse, setzen voraus, dass sich die zu analysierenden Daten in einer einzigen Spalte des Datenblatts befinden, zusammen mit einer oder mehreren Kodespalten, welche die erklärenden Faktoren identifizieren. Nehmen Sie die Daten in der folgenden Kreuztabelle als Beispiel:

Mischung	*Behandlung 1*	*Behandlung 2*	*Behandlung 3*
1	75	82	91
2	78	85	93
3	77	84	92
4	75	85	96

Um diese Daten mit der Prozedur *Mehrfaktorielle ANOVA* zu analysieren, werden sie folgendermaßen in ein Datenblatt eingefügt:

<unbezeichnet>

	Mischung	Behandlung	Y	Col_4	Col_5	Col_6	Col_7
1	1	1	75				
2	1	2	82				
3	1	3	91				
4	2	1	78				
5	2	2	85				
6	2	3	93				
7	3	1	77				
8	3	2	84				
9	3	3	92				
10	4	1	75				
11	4	2	85				
12	4	3	96				
13							
14							
15							
16							
17							
18							

A / B / C / D / E / F / G / H / I / J

Abbildung 2.23: Gewünschte Datenstruktur

Die ersten zwei Spalten zeigen die Faktorstufen zu den entsprechenden Datenwerten an. Die dritte Spalte enthält die Beobachtungswerte.

Um eine solche Datei zu erstellen, ist es häufig die einfachste Lösung, die ersten beiden Spalten manuell einzugeben. Da die Spalten aber einfachen Mustern folgen, lassen sie sich auch mit

Hilfe bestimmter STATGRAPHICS Centurion-Operatoren generieren. Die Zahlen für *Mischung* können z. B. generiert werden, indem Sie die Spalte 1 markieren und im *Bearbeiten*-Menü *Daten generieren* auswählen. Sie erhalten die folgende Dialogbox, in die der Ausdruck bereits eingegeben wurde:



Abbildung 2.24: Generieren der Zahlen für Mischung

Die Option *Daten generieren* wertet den STATGRAPHICS Centurion-Ausdruck aus und fügt das Ergebnis in die gewählte Spalte ein. In dem obigen Ausdruck werden zwei wichtige Operatoren angewendet:

> *COUNT(von;bis;Schrittweite)* – generiert die Werte, angefangen bei *von*, endend bei *bis* und in Intervallen, die *Schritweite* entsprechen. COUNT(1;4;1) generiert also die ganzen Zahlen 1, 2, 3 und 4.
>
> *REP(X;Wiederholungen)* – wiederholt jeden Wert von *X* in Gruppen so oft wie bei *Wiederholungen* angegeben. Im vorliegenden Fall wird jede ganze Zahl zwischen 1 und 4 3-mal wiederholt.

Die Zahlen für *Behandlung* können auf eine ähnliche Weise generiert werden, indem Sie den Spaltenkopf der Spalte 2 markieren, im *Bearbeiten*-Menü *Daten generieren* auswählen und Folgendes in die aufgerufene Dialogbox eingeben:

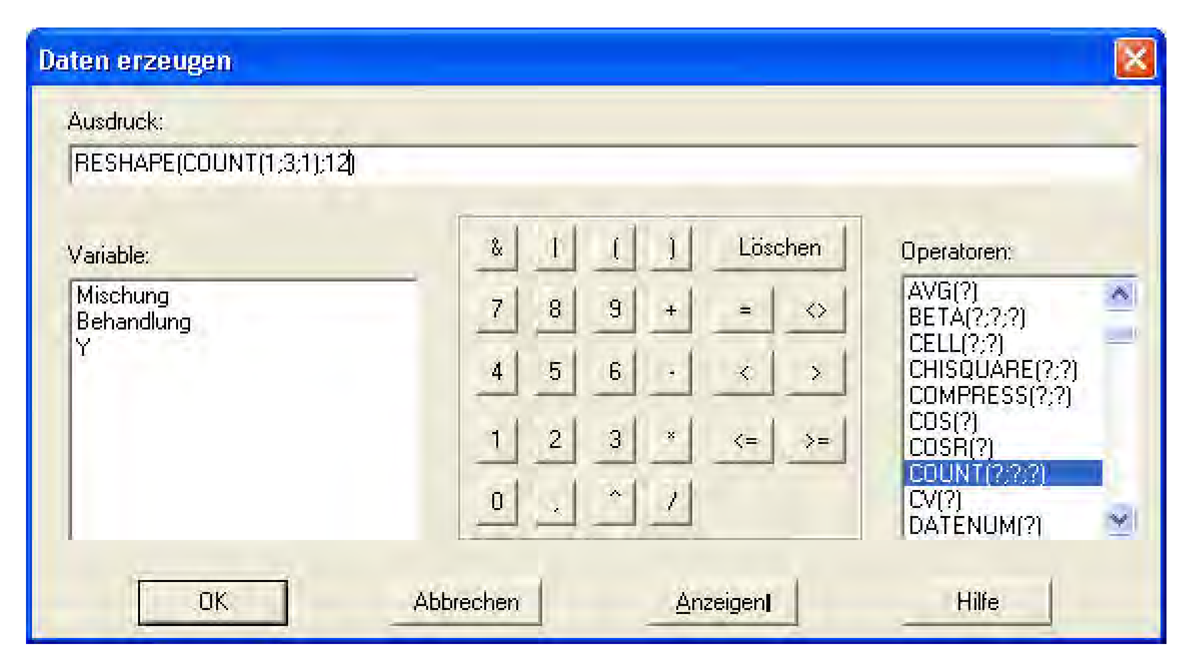


Abbildung 2.25: Generieren der Zahlen für Behandlung

Der Ausdruck verwendet einen weiteren Operator:

> *RESHAPE(X;Größe)* – wiederholt periodisch die Werte in *X,* bis so viele Werte, wie bei *Größe* angegeben, generiert sind. Im vorliegenden Fall wird die Folge 1, 2, 3 insgesamt 4-mal wiederholt.

Die Struktur-Generatoren sind besonders bei der Erstellung großer Datendateien hilfreich.

2.4.2 Generieren von Zufallszahlen

Zufallszahlen können mit STATGRAPHICS Centurion XVI auf zwei Arten generiert werden:

1. Stammen die Zahlen aus einer Exponential-, Gamma-, Lognormal-, Normal-, Gleich- oder Weibull-Verteilung, so können sie im Datenblatt durch Markieren des Spaltenkopfs, anschließender Wahl von *Daten generieren* im *Bearbeiten*-Menü und Eingabe des entsprechenden STATGRAPHICS Centurion-Ausdrucks generiert werden.

2. Bei anderen Verteilungen werden die Zufallszahlen über die *Wahrscheinlichkeitsverteilung*-Prozedur generiert.

Angenommen, es würden 100 Zufallszahlen aus einer Normalverteilung mit einem Mittelwert von 20 und einer Standardabweichung von 2 benötigt. Markieren Sie den Kopf einer leeren Spalte des Datenblatts. Wählen Sie dann *Daten generieren* im *Bearbeiten*-Menü und befüllen Sie die aufgerufene Dialogbox wie folgt:

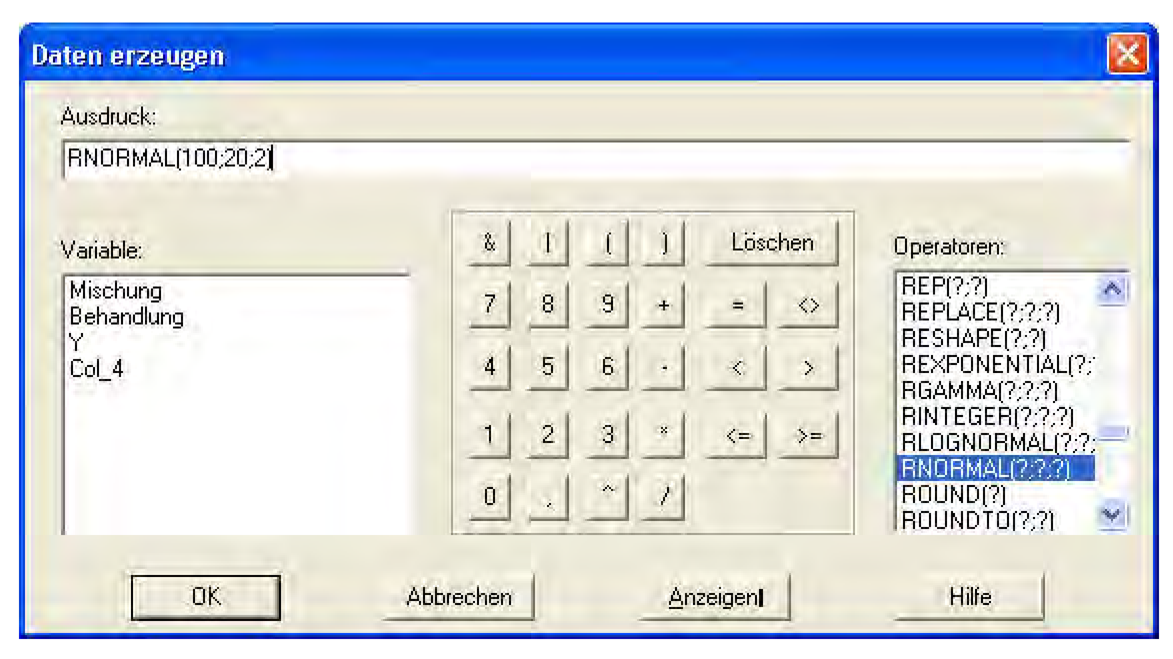


Abbildung 2.26: Generieren von Zufallszahlen aus einer Normalverteilung

Die Syntax des RNORMAL-Operators besteht aus:

> *RNORMAL*(*n*;*mu*;*sigma*) – generiert *n* Pseudo-Zufallszahlen aus einer Normalverteilung mit dem Mittelwert *mu* und der Standardabweichung *sigma*.

Wählen Sie *OK*, um die Zufallszahlen zu generieren und sie in der markierten Spalte einzufügen.

Die Syntax weiterer Zufallszahlen-Generatoren wird im PDF-Dokument *STATGRAPHICS Centurion Operators* beschrieben.

2.5 Die Eigenschaften des Datenbuchs

Die vorhergehenden Abschnitte befassten sich mit den wichtigsten Aspekten des Datenhandlings in STATGRAPHICS Centurion XVI. Insbesondere wurde gezeigt, wie Daten aus Dateien und Datenbanken gelesen und wie sie anschließend im STATGRAPHICS

Centurion XVI-Datenblatt weiter bearbeitet werden können. Der Status eines Datenblatts kann zu jeder beliebigen Zeit abgerufen werden. Hierfür aktivieren Sie das *Datenbuch*-Fenster und wählen *Datenbuch-Eigenschaften* im *Bearbeiten*-Menü aus oder, alternativ, *StatLink* im *Datei*-Menü:

Abbildung 2.27: Dialogbox mit den Datenbuch-Eigenschaften

Die aufgerufene Dialogbox führt die aktuellen Datenquellen für jedes Datenblatt auf. Sofern erwünscht, können die Datenblätter schreibgeschützt werden, um zu verhindern, dass die enthaltenen Daten versehentlich geändert werden. Zudem kann hier eingestellt werden, dass die Datenquellen in regelmäßigen Abständen abgefragt und statistische Prozeduren automatisch aktualisiert werden. Diese und weitere wichtige Funktionalitäten werden in Kapitel 5 beschrieben.

2.6 Daten-Betrachter

STATGRAPHICS Centurion XVI enthält eine neue Prozedur, um Daten zu betrachten. Die Prozedur – die Sie über *Werkzeuge – Datenbetrachter* aufrufen können – erzeugt eine Gesamtübersicht über die Anzahlen der nicht-fehlenden und eindeutigen Werte sowie der Minimum- und Maximumwerte der ausgewählten Variablen:

Daten-Betrachter
Anzahl Spalten: 28
Anzahl Zeilen: 100
Anzahl vollständiger Fälle: 82

Spalte			*Nichtfehlende*	*Eindeutige*		
Name	*Kommentar*	*Typ*	*Werte*	*Werte*	*Minimum*	*Maximum*
Make		Text	93	32		
Model		Text	93	93		
Type		Text	93	6		
Min Price	price for basic version in $1,000	Numerisch	93	79	6,7	45,4
Mid Price	average of min and max prices in $1,000	Numerisch	93	81	7,4	61,9
Max Price	price for a premium version in $1,000	Numerisch	93	79	7,9	80,0
MPG City	miles per gallon in city driving	Numerisch	93	21	15,0	46,0
MPG Highway	miles per gallon in highway driving	Numerisch	93	22	20,0	50,0
Air Bags	0=none, 1=driver only, 2=driver and passenger	Numerisch	93	3	0	2,0
Drive Train		Text	93	3		
Cylinders		Numerisch	92	5	3,0	8,0
Engine Size	liters	Numerisch	93	26	1,0	5,7
Horsepower	maximum	Numerisch	93	57	55,0	300,0
RPM	revs per minute at maximum horsepower	Numerisch	93	24	3800,0	6500,0
Revs per Mile	revs per mile in highest gear	Numerisch	93	78	1320,0	3755,0
Manual	0=no, 1=yes	Numerisch	93	2	0	1,0
Fueltank	gallons	Numerisch	93	38	9,2	27,0
Passengers	persons	Numerisch	93	6	2,0	8,0
Length	inches	Numerisch	93	51	141,0	219,0
Wheelbase	inches	Numerisch	93	27	90,0	119,0
Width	inches	Numerisch	93	16	60,0	78,0
U Turn Space	feet	Numerisch	93	14	32,0	45,0
Rear seat	inches	Numerisch	91	24	19,0	36,0
Luggage	cu. ft.	Numerisch	82	16	6,0	22,0
Weight	pounds	Numerisch	93	81	1695,0	4105,0
Domestic	1=U.S. manufacturer	Numerisch	93	2	0	1,0
Date		Datum	100	100	20090,0	20189,0
Strength		Numerisch	100	77	191,3	229,5

Abbildung 2.28: Ausgabe des Daten-Betrachters

Chapter

3

Statistische Analysen

Generieren von Analysen, Auswählen der Tabellen und Grafiken, Auswahl von Optionen, Verändern der Eingabedaten und Sichern der Ergebnisse.

Das STATGRAPHICS Centurion XVI-Hauptmenü umfasst mehr als 160 statistische Prozeduren. Hinter jeder Auswahl verbirgt sich jeweils eine andere statistische Prozedur. Die Prozeduren folgen einem einheitlichen Ablaufschema:

1. Mit der Auswahl einer Analyse aus dem Hauptmenü wird eine *Dateneingabe-Dialogbox* geöffnet. Die Felder in der Dialogbox dienen dazu, die zu analysierenden Variablen festzulegen.

2. Wenn die gewählte Prozedur Optionen umfasst, die alle Tabellen und Diagramme der Prozedur beeinflussen, wird eine *Analyse-Optionen*-Dialogbox angezeigt, über die Sie die Optionen der Prozedur modifizieren können.

3. Wenn die gewählte Prozedur mehr als eine einzige Tabelle und ein einziges Diagramm umfasst, wird die *Tabellen und Grafiken*-Dialogbox angezeigt, über die Sie die benötigten Tabellen und Diagramme auswählen.

4. Die festgelegten Daten werden nun gelesen und analysiert, und ein neues *Analyse-Fenster* wird erstellt.

5. Die vorher gewählten Optionen können über die Schaltfläche *Analyse-Optionen* in der Analyse-Symbolleiste angepasst werden. Die von der Anpassung betroffenen Tabellen und Diagramme werden automatisch aktualisiert.

6. Weitere Tabellen und Diagramme können über die Schaltfläche *Tabellen und Grafiken* auf der Analyse-Symbolleiste aufgerufen werden.

7. Die einzelnen Tabellen und Grafiken können durch Maximieren des entsprechenden Ergebnisfensters und anschließendes Aufrufen der *Ergebnisfenster-Optionen* auf der Analyse-Symbolleiste angepasst werden.

8. Die Standardwerte für Diagramme, wie Titel, Skalierung, Schriften und Zeichen, können im maximierten Ergebnisfenster über die Schaltfläche *Diagramm-Optionen* auf der Analyse-Symbolleiste angepasst werden.

9. Tabellen und Diagramme können gedruckt, als HTML-Dateien publiziert, in andere Anwendungen wie Microsoft PowerPoint kopiert oder im StatReporter gesichert werden.

10. Numerische Ergebnisse können über die Schaltfläche *Ergebnisse speichern* auf der Analyse-Symbolleiste in einer beliebigen Spalte eines Datenblatts gesichert werden.

11. Die vollständige Analyse kann in Form eines *StatFolio* gesichert werden, das zu einem späteren Zeitpunkt weiter bearbeitet werden kann.

Dieses Kapitel beschreibt eine typische Analyse im Detail. Das Ziel der Analyse ist, ein statistisches Modell zu entwickeln, das die Beziehung zwischen der Reichweite *MPG City* für die n = 93 Fahrzeuge aus der *93cars.sgd*-Datei und ihrem Gewicht *Weight* wiedergibt. Das folgende Streudiagramm beinhaltet die entsprechenden Daten:

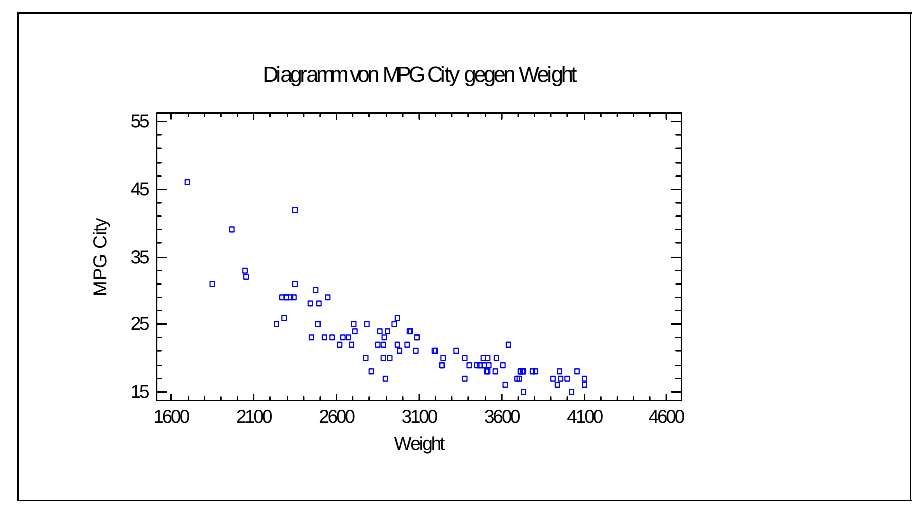


Abbildung 3.1: x-y-Diagramm für MPG City und das Gewicht

Wie Sie vermutlich erwartet haben, steht die Reichweite *MPG City* in negativer Korrelation zum Fahrzeuggewicht. In der Beziehung ist eine gewisse Nichtlinearität erkennbar; zudem verhält sich mindestens ein Punkt wie ein Ausreißer.

Die grundlegende Prozedur in STATGRAPHICS Centurion XVI zur Anpassung eines statistischen Modells an zwei Variablen ist die *Einfache Regresssion*. Sie kann für lineare und nicht-lineare Modelle angewendet werden. Das einfachste Modell, das eine abhängige Variable Y zu einer unabhängigen Variable X in Beziehung setzt, ist eine Gerade der Form

$$Y = a + b\,X$$

mit b gleich der Steigung der Geraden und a gleich dem y-Achsenabschnitt. Kurvenlineare Modelle wie das Exponentialmodell

$$Y = exp(a + b\,X)$$

können für nicht-lineare Beziehungen verwendet werden.

3.1 Die Dateneingabe-Dialogboxen

Die Prozedur *Einfache Regression* befindet sich im Hauptmenü an folgender Stelle:

1. Im klassischen Menü unter *Beziehungen – Ein Faktor.*

2. Im Six Sigma-Menü unter *Verbessern – Regressionsanalyse – Ein Faktor.*

Mit dem Beginn der Prozedur wird folgende typische Dateneingabe-Dialogbox angezeigt:



Abbildung 3.2: Dateneingabe-Dialogbox für die Einfache Regression

Die ersten beiden Eingabefelder müssen ausgefüllt werden:

Y: Die abhängige Variable.

X: Die unabhängige Variable.

Sie können hier entweder die Spaltennamen wie *MPG City* oder einen STATGRAPHICS Centurion-Ausdruck wie *LOG(MPG City)* eingeben. Wenn mehr als ein Datenblatt des Datenbuchs den angegebenen Spaltennamen enthält, geben Sie vor dem Spaltennamen das gewünschte Datenblatt an. Wenn z. B. die Datenblätter *A* und *B* eine Spalte mit dem Namen *Weight* enthalten und Sie die Spalte des Datenblatts *A* verwenden möchten, tragen Sie *A.Weight* in das Feld ein.

In das Eingabefeld *Bedingung* kann der auszuwertende Bereich an Zeilen des Datenblatts eingetragen werden. Wenn Sie hier z. B. eine Anweisung wie *FIRST(50)* eingeben, werden nur die ersten 50 Zeilen des Datenblatts in die Analyse einbezogen. Typische Angaben für das Feld *Bedingung* sind:

<table>
<tr><th>Eingabe</th><th>Aktion</th><th>Beispiel</th></tr>
<tr><td>FIRST(k)</td><td>Wählt die ersten k Zeilen.</td><td>FIRST(50)</td></tr>
<tr><td>LAST(k)</td><td>Wählt die letzten k Zeilen.</td><td>LAST(50)</td></tr>
<tr><td>ROWS(Start;Ende)</td><td>Wählt alle Zeilen zwischen und einschließlich Start und Ende.</td><td>ROWS(21;70)</td></tr>
<tr><td>RANDOM(k)</td><td>Wählt k Zeilen zufällig aus.</td><td>RANDOM(50)</td></tr>
<tr><td>Spaltenname < Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist kleiner als Wert .</td><td>Passengers < 5</td></tr>
<tr><td>Spaltenname <= Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist kleiner als oder gleich Wert .</td><td>Passengers <= 5</td></tr>
<tr><td>Spaltenname> Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist größer als Wert .</td><td>Passengers > 5</td></tr>
<tr><td>Spaltenname>= Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist größer als oder gleich Wert .</td><td>Passengers >= 5</td></tr>
<tr><td>Spaltenname= Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist gleich Wert .</td><td>Cylinders = 6</td></tr>
<tr><td>Spaltenname<> Wert</td><td>Wählt nur die Zeilen mit Spaltenname ist nicht gleich Wert .</td><td>Cylinders <> 4</td></tr>
<tr><td>Bedingung1 & Bedingung2</td><td>Wählt nur die Zeilen aus, die beide Bedingungen erfüllen.</td><td>Cylinders = 6 & Make = “Ford”</td></tr>
<tr><td>Bedingung1 | Bedingung2</td><td>Wählt nur die Zeilen aus, die wenigstens eine der Bedingungen erfüllen.</td><td>Cylinders = 6 | Make = “Ford”</td></tr>
<tr><td>Binärspalte</td><td>Wählt nur die Zeilen aus, deren Werte in Binärspalte nicht gleich 0 sind.</td><td>Domestic</td></tr>
</table>

Abbildung 3.3: Mögliche Eingaben für das Eingabefeld Bedingung

Wenn Sie eine Bedingung angeben, die eine nicht-numerische Variable enthält, muss *Wert* in Anführungszeichen gesetzt werden. Bei *Wert* wird zwischen Groß- und Kleinbuchstaben unterschieden. Mehrere Bedingungen können mit dem logischen UND (&) bzw. ODER (|) verknüpft werden.

Jede der möglichen Eingaben für das Feld *Bedingung* generiert Boolsche Werte, wobei *0* für *FALSCH* und *1* für *WAHR* steht. In der Analyse werden dann alle Zeilen berücksichtigt, für welche die Bedingung *WAHR* ist, und alle Zeilen ausgeschlossen, für welche die Bedingung *FALSCH* ist.

3.2 Das Analyse-Fenster

Nach dem Festlegen der Daten wird ein neues *Analyse-Fenster* erstellt:

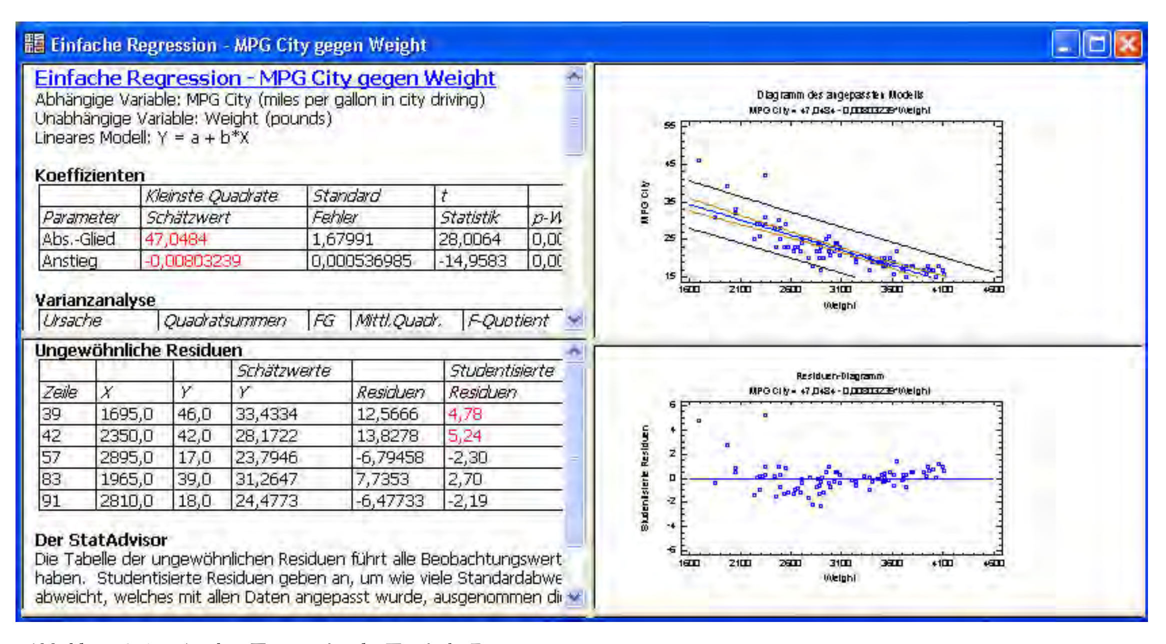


Abbildung 3.4: Analyse-Fenster für die Einfache Regression

Das Fenster ist durch bewegliche Balken in mehrere *Ergebnisfenster* unterteilt. Die Tabellen der Analyse befinden sich im linken Bereich des Fensters, während die Diagramme im rechten Bereich angezeigt werden.

Sie können eine Tabelle oder ein Diagramm vergrößern, wenn Sie mit der linken Maustaste doppelt in das entsprechende Ergebnisfenster klicken:

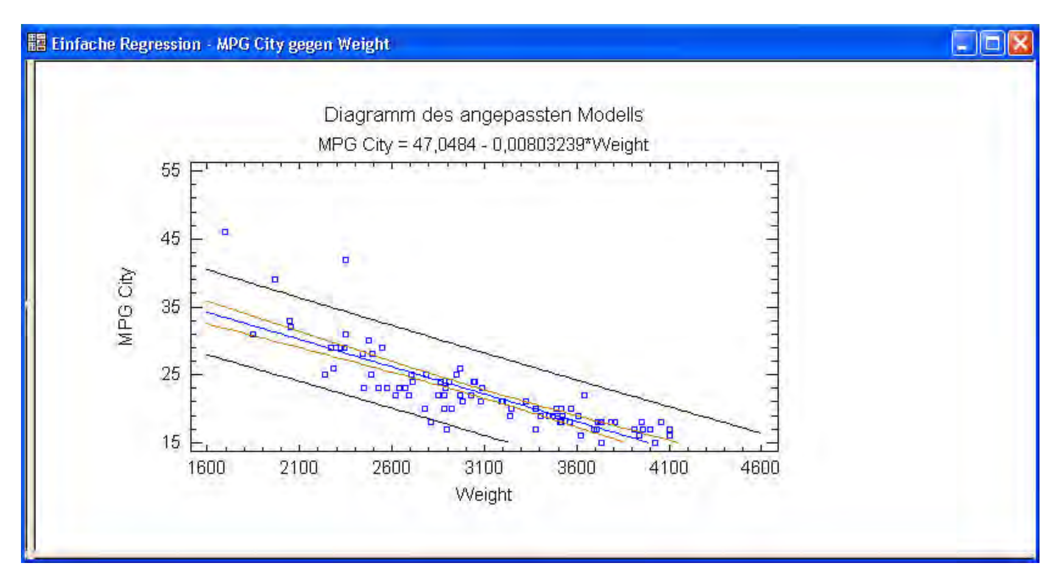
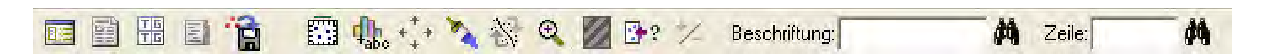


Abbildung 3.5: Maximiertes Ergebnisfenster aus dem Analyse-Fenster für die Einfache Regression

Über einen Doppelklick in das Ergebnisfenster gelangen Sie zurück in die vorhergegangene Anzeige.

Ist das Analyse-Fenster im Vordergrund, so wird neben der STATGRAPHICS Centurion XVI-Symbolleiste eine weitere Symbolleiste aktiviert. Diese *Analyse-Symbolleiste* sieht wie folgt aus:



Jeder der Schaltflächen auf der Symbolleiste liegt eine wichtige Aktion zugrunde.

3.2.1 Die Schaltfläche *Dateneingabe*

Wenn Sie diese Schaltfläche auswählen, wird die Dateneingabe-Dialogbox angezeigt, die ursprünglich für die Festlegung der Datenvariablen verwendet wurde (siehe *Abbildung 3.2*). Sie können nun die Datenvariablen ändern. Wählen Sie *OK*, um die Änderungen auf Ihre Analyse anzuwenden. Dies ermöglicht das Arbeiten mit verschiedenen Datenkombinationen, ohne eine neue Analyse zu starten.

3.2.2 Die Schaltfläche *Analyse-Optionen*

Für die meisten der Analysen stehen zahlreiche Optionen bereit. Bei der erstmaligen Durchführung einer Analyse werden Standardoptionen verwendet, die in vielen Fällen auch genügend Information ausgeben. Über die Schaltfläche *Analyse-Optionen* können Sie die Standardeinstellungen für jede Prozedur ändern. Für die *Einfache Regression* kann in der Dialogbox der *Analyse-Optionen* der Typ des angepassten Modells und die Methode für die Schätzung der unbekannten Modellkoeffizienten festgelegt werden:

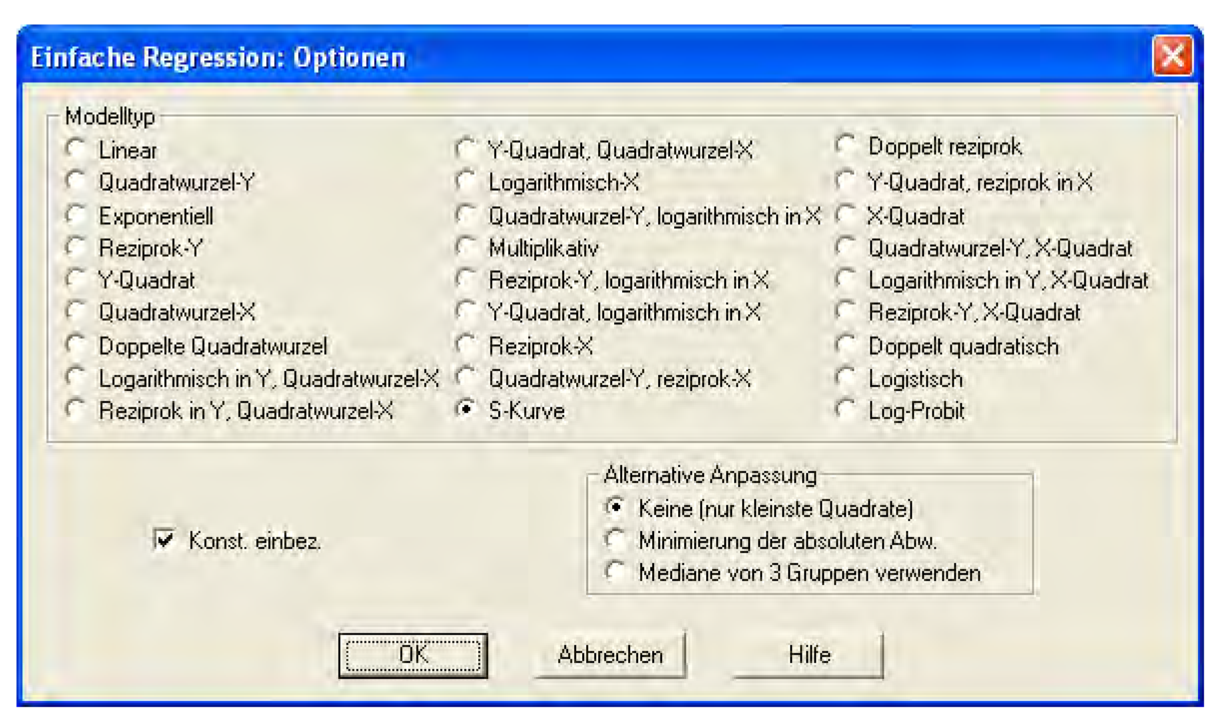


Abbildung 3.6: Dialogbox mit den Analyse-Optionen für die Einfache Regression

Wenn Sie die Ergebnisse in *Abbildung 3.9* betrachten, sehen Sie, dass in der Tabelle mit dem Vergleich der alternativen Modelle die kurvilinearen Modelle größere R-Quadrat-Werte besitzen als das lineare Modell. An erster Stelle der Auflistung steht das *S-Kurven*-Modell. Wenn Sie dieses Modell in der Dialogbox *Analyse-Optionen* markieren und *OK* wählen, wird die gesamte Auswertung so geändert, dass sie das neue Modell wiedergibt. Wie in dem Diagramm des angepassten Modells zu erkennen ist, kann mit der S-Kurve die Krümmung der Daten sehr gut dargestellt werden:

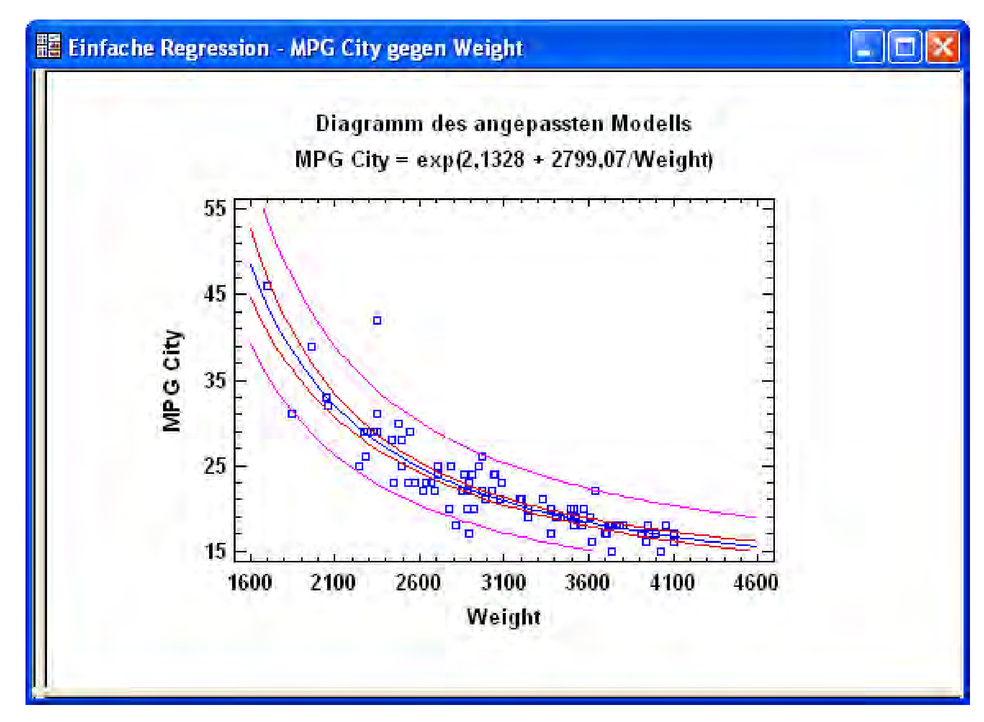


Abbildung 3.7: Angepasstes S-Kurven-Modell

3.2.3 Die Schaltfläche *Tabellen und Grafiken*

Wenn Sie diese Schaltfläche auswählen, wird eine Liste mit den Tabellen und Diagrammen angezeigt, die in das Analyse-Fenster eingefügt werden können. Für die *Einfache Regression* sind folgende Tabellen und Diagramme verfügbar:

Tabellen und Grafiken

TABELLEN
- [x] Zusammenfassung der Analyse
- [ ] Lack-of-Fit-Test
- [ ] Vorhersagen
- [x] Vergleich der alternativen Modelle
- [x] Ungewöhnliche Residuen
- [ ] Einflussreiche Punkte

GRAFIKEN
- [x] Diagramm des angepassten Modells
- [ ] Beobachtungswerte gegen Modellwerte
- [x] Residuen gegen X
- [ ] Residuen gegen Modellwerte
- [ ] Residuen gegen Zeilennummer

OK
Abbrechen
Alles
uswahl speiche
Hilfe

Abbildung 3.8: Dialogbox mit den für die Einfache Regression verfügbaren Tabellen und Grafiken

Wenn Sie z.B. auswählen, dass Tabellen mit alternativen Modellen und ungewöhnlichen Residuen angezeigt werden sollen, so werden diese neuen Text-Ergebnisfenster dem Analyse-Fenster hinzugefügt:

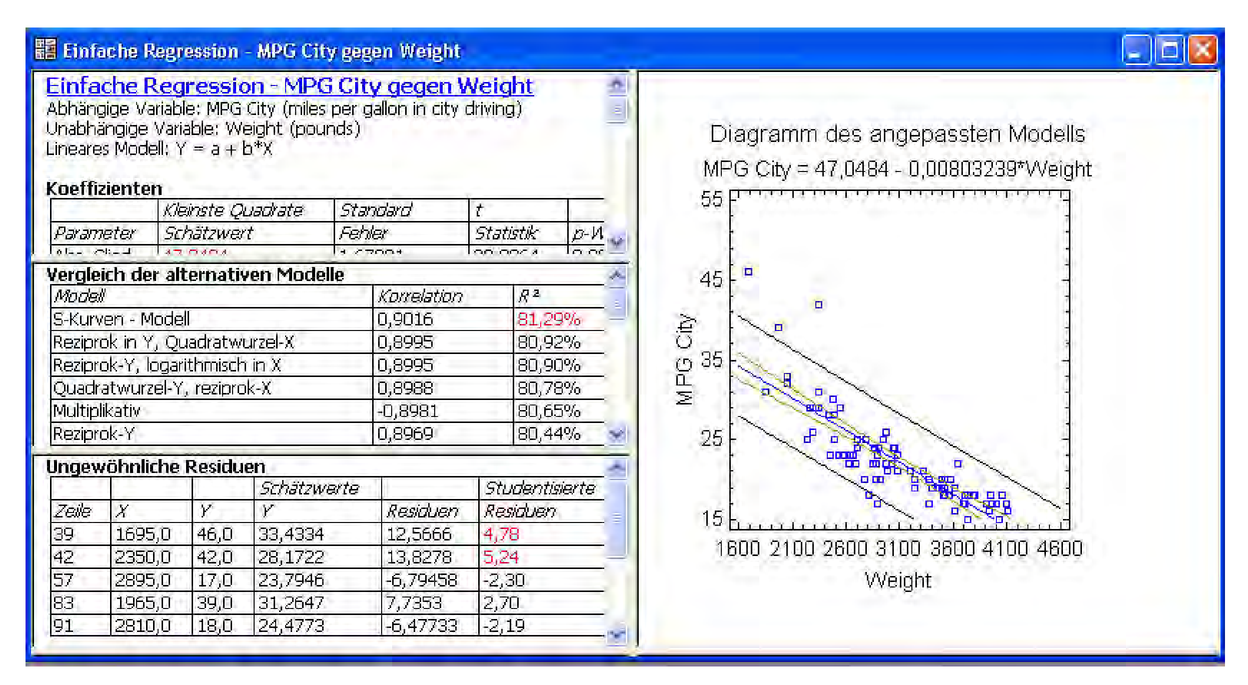


Abbildung 3.9: Analyse-Fenster der Einfachen Regression mit den eingefügten Tabellen

Sie können dem Analyse-Fenster auch ein weiteres Diagramm, z. B. ein Residuen-Diagramm hinzufügen:

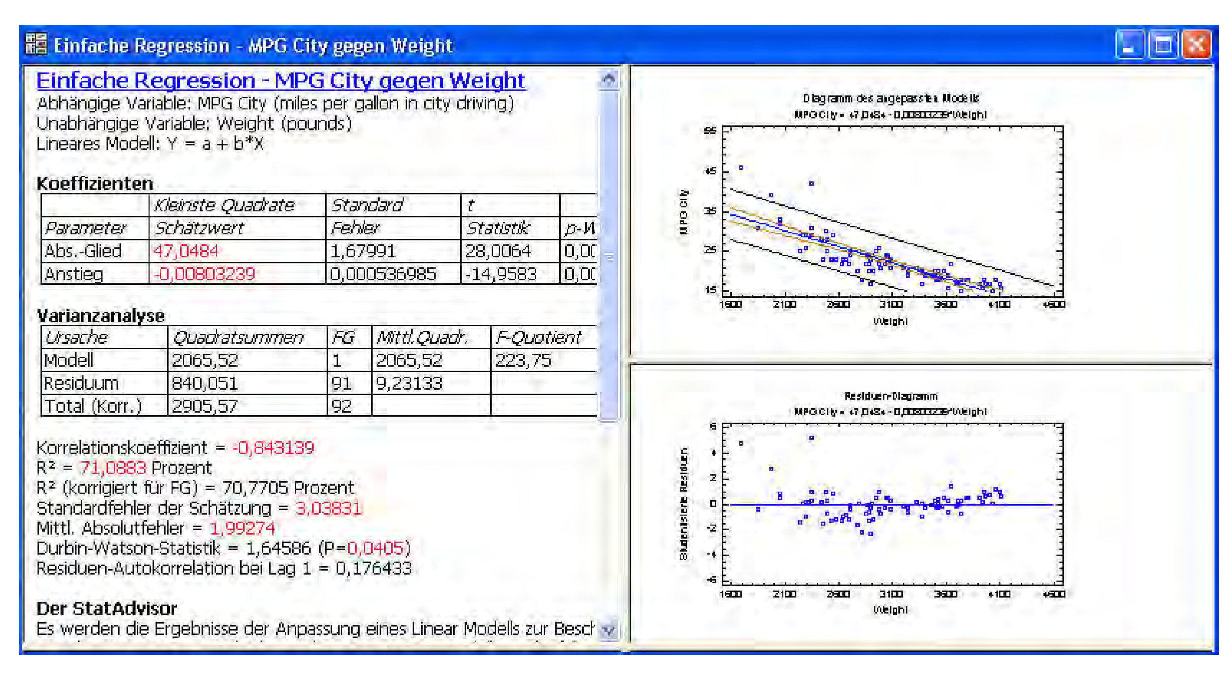


Abbildung 3.10: Analyse-Fenster der Einfachen Regression mit dem zusätzlichen Diagramm

3.2.4 Die Schaltfläche *Ergebnisfenster-Optionen*

Zusätzlich zu den Optionen, die für das gesamte Analyse-Fenster verfügbar sind, gibt es zahlreiche Tabellen und Diagramme mit eigenen Optionen. Auf diese Optionen können Sie zugreifen, indem Sie die gewünschte Tabelle oder Grafik maximieren und dann die Schaltfläche *Ergebnisfenster-Optionen* wählen. Für das *Diagramm des angepassten Modells* können folgende Ergebnisfenster-Optionen festgelegt werden:



Abbildung 3.11: Dialogbox mit den Ergebnisfenster-Optionen für das Diagramm des angepassten Modells

Wenn Sie nun z.B. das Kontrollkästchen *Konfidenzgrenzen* deaktivieren und *OK* klicken, wird das Diagramm ohne die inneren Grenzen neu erstellt:

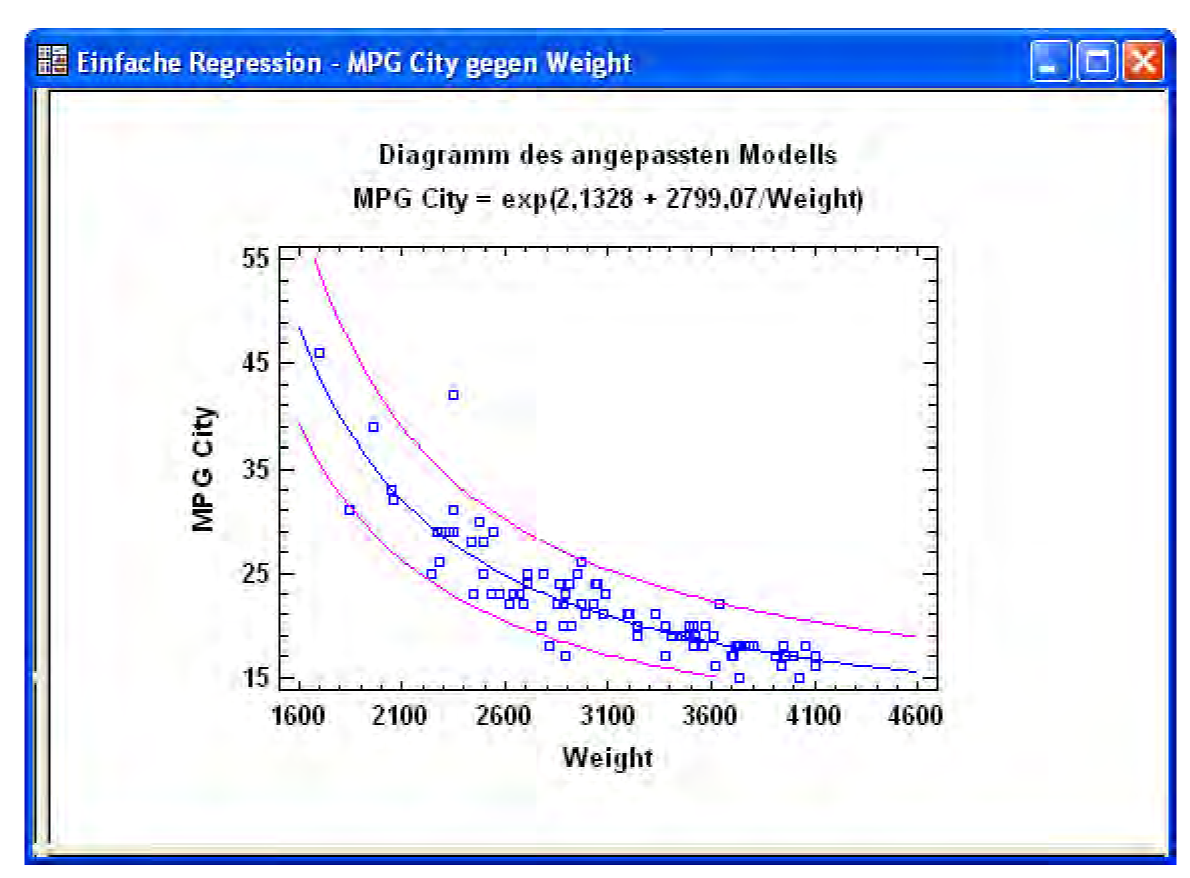


Abbildung 3.12: Diagramm des angepassten Modells ohne Konfidenzgrenzen

3.2.5 Die Schaltfläche *Ergebnisse speichern*

Mit dieser Schaltfläche können Sie die numerischen Ergebnisse, die in einer statistischen Analyse berechnet wurden, in den Spalten eines Datenblatts sichern. Für die *Einfache Regression* werden die folgenden Speicheroptionen angeboten:

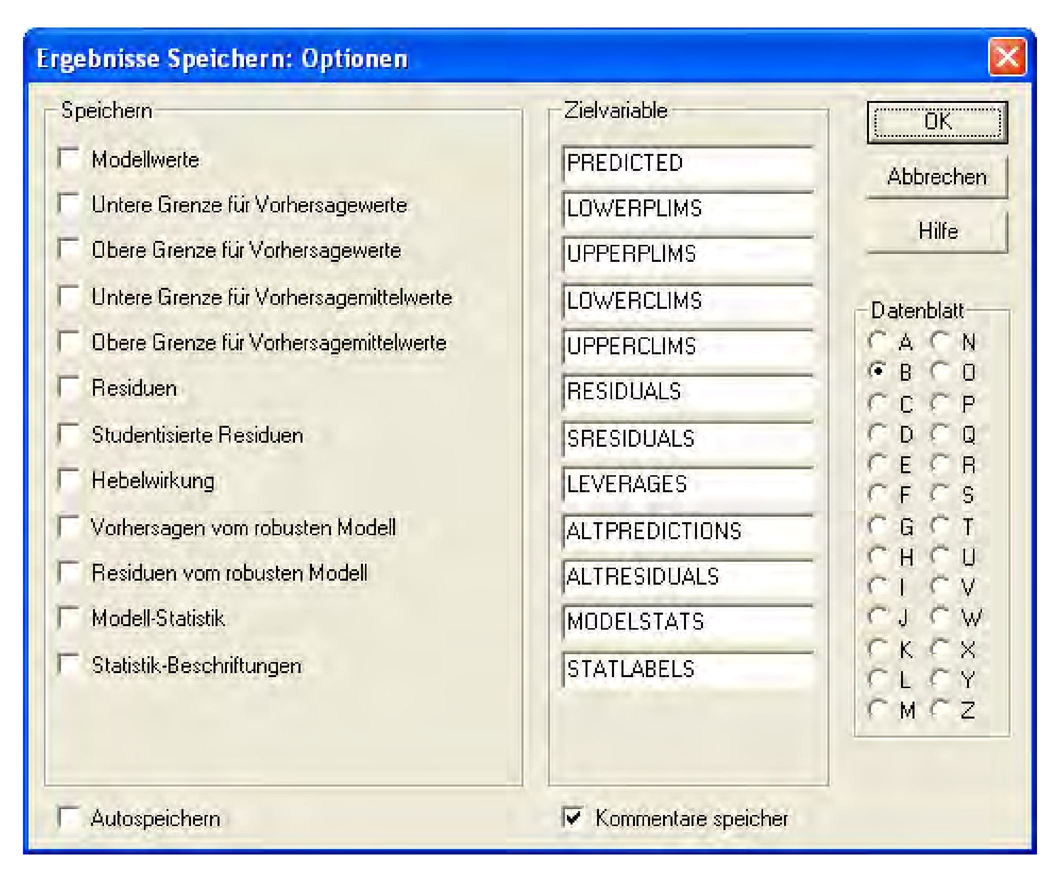


Abbildung 3.13: Dialogbox zum Speichern der numerischen Ergebnisse der Einfachen Regression

Um die gewünschten Informationen zu sichern, markieren Sie die passenden Kontrollkästchen im Gruppenfeld *Speichern*. Für die markierten, d.h. die zu sichernden Informationen geben Sie im Gruppenfeld *Zielvariable* einen Spaltennamen und im Gruppenfeld *Datenblatt* das Datenblatt an. Möchten Sie zusätzlich zu den Daten auch die Kommentare sichern, dann markieren Sie das Kontrollkästchen *Kommentare speichern*.

Das Kontrollkästchen *Autospeichern* kann markiert werden, wenn die zu sichernden Informationen bei jedem Analysevorgang automatisch gespeichert werden sollen. Dies bietet

sich an, wenn Sie Ihre Auswertung als StatFolio speichern, da die Analysen beim Öffnen eines StatFolios prinzipiell aktualisiert werden. Mit *Autospeichern* können Sie also ein StatFolio erzeugen, das die ausgewählten Statistiken automatisch errechnet und speichert. In Kombination mit dem Schreiben eines *Scripts* , was in Kapitel 5 beschrieben wird, ermöglicht diese Funktion, viele Aufgaben und Aktionen zu automatisieren.

3.2.6 Schaltflächen für Grafiken

Wann immer Sie ein Diagramm in einem Analyse-Fenster maximieren, werden einige zusätzliche Schaltflächen wie z.B. die nachfolgenden aktiviert:

Diagramm-Optionen – zeigt eine Dialogbox an, über die Farbe, Beschriftungen, Achsenskalierung und ähnliche Eigenschaften geändert werden können.

Text hinzufügen – fügt dem Diagramm zusätzlichen Text hinzu.

Auseinanderziehen – versetzt die Datenpunkte eines Diagramms zufällig in horizontaler oder vertikaler Richtung, um zu verhindern, dass sie sich überschneiden.

Einfärben – färbt die Datenpunkte eines Streudiagramms mit einer Farbe ein entsprechend dem Wert einer gewählten Variable.

Glätten/Rotieren – glättet ein zweidimensionales Diagramm bzw. dreht ein dreidimensionales Diagramm.

Zoomen und Schwenken – schwenkt und zoomt das Diagramm entlang der X-, Y- oder Z-Achse.

Untersuchen – dynamisches Untersuchen von Flächen- und Konturen-Diagrammen

Identifizieren – zeigt eine Beschriftung mit den Informationen zu einem Punkt an, wenn dieser mit der linken Maustaste angeklickt wird.

Lokalisieren mit Namen – zeigt diejenigen Datenpunkte rot an, deren Werte den im Feld *Beschriftung* eingegebenen Werten entsprechen (wird zusammen mit der Schaltfläche *Identifizieren* verwendet).

Lokalisieren mit Zeilennummer – zeigt diejenigen Datenpunkte rot an, deren Werte den im Feld *Zeile* eingegebenen Zeilennummern entsprechen.

In Kapitel 4 werden diese Schaltflächen ausführlicher beschrieben.

3.2.7 Die Schaltfläche *Entfernen*

In einigen der statistischen Analysen können potenzielle Ausreißer von der Auswertung ausgeschlossen werden. Hierzu maximieren Sie das Diagramm, markieren den entsprechenden Punkt und wählen die Schaltfläche *Entfernen* aus. Das Diagramm in *Abbildung 3.12* zeigt z.B. einen Punkt, der sich deutlich außerhalb der Vorhersagegrenzen befindet. Wenn Sie den Punkt mit der Maus anklicken und *Entfernen* wählen, wird das Modell ohne den entsprechenden Punkt angepasst. Das Diagramm des angepassten Modells zeigt dann ein neues Modell, in dem der entfernte Punkt (oder Punkte) durch ein *X* ersetzt ist.

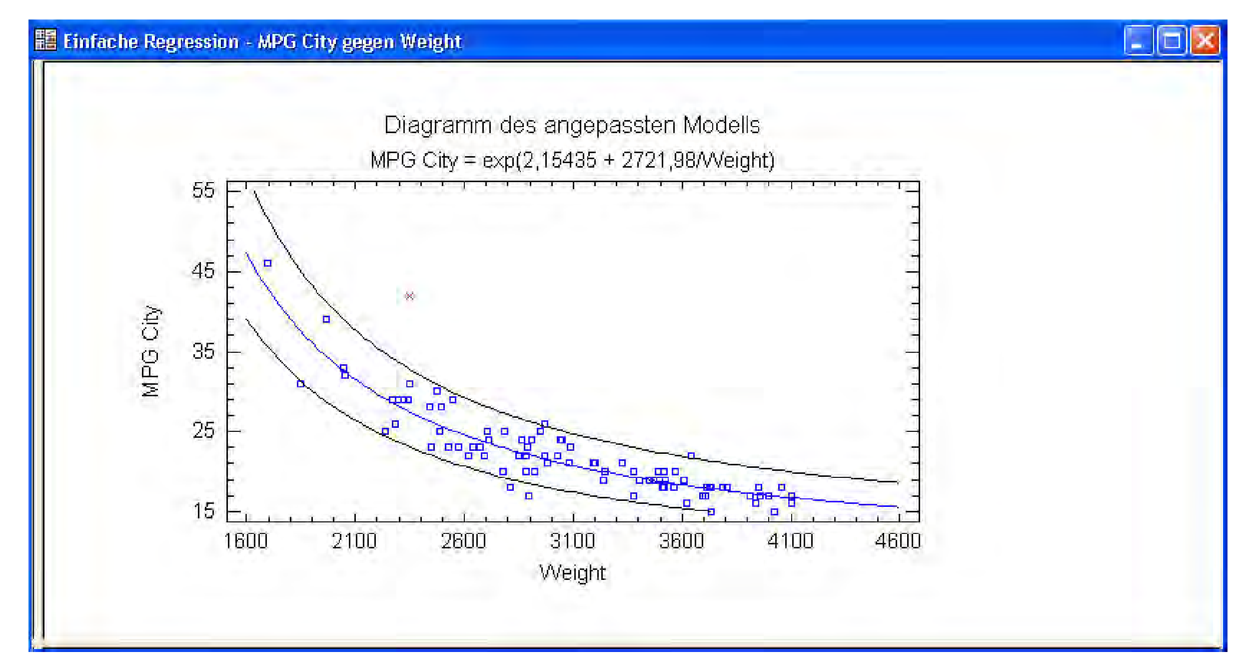


Abbildung 3.14: Angepasstes S-Kurven-Modell nach dem Ausschluss eines möglichen Ausreißers

Zudem werden auch alle weiteren Tabellen und Diagramme im Analyse-Fenster an die Veränderung des Modells angepasst.

Mehrere Punkte können aus einem Modell ausgeschlossen werden, indem sie nacheinander markiert und mit *Entfernen* entfernt werden. Um einen entfernten Punkt wieder in das Modell einzubeziehen, wird auf ihn mit der linken Maustaste und dann erneut auf *Entfernen* geklickt.

3.3 Drucken der Ergebnisse

Die Ergebnisse einer statistischen Analyse können ganzheitlich oder einzeln gedruckt werden:

1. Um alle Tabellen und Diagramme des Analyse-Fensters zu drucken, wählen Sie die Schaltfläche *Drucken* auf der Analyse-Symbolleiste oder wählen *Drucken* im *Datei*-Menü

2. Um eine einzige Tabelle oder Diagramm zu drucken, rufen Sie das Kontextmenü des entsprechenden Ergebnisfensters auf (rechter Mausklick) und wählen *Drucken*.

Die folgende Dialogbox wird angezeigt, wenn Sie die vollständige Analyse drucken möchten:



Abbildung 3.15: Dialogbox mit den Druckoptionen für eine vollständige Analyse

In dem Gruppenfeld *Druckbereich* bestimmen Sie, welche Ergebnisfenster gedruckt werden sollen. Sie können das Kontrollkästchen *Alle Analysen* wählen, um die Ergebnisse aller Analyse-Fenster zu drucken.

Weitere Optionen für den Druck können Sie über die Option *Seite einrichten* im *Datei*-Menü festlegen:

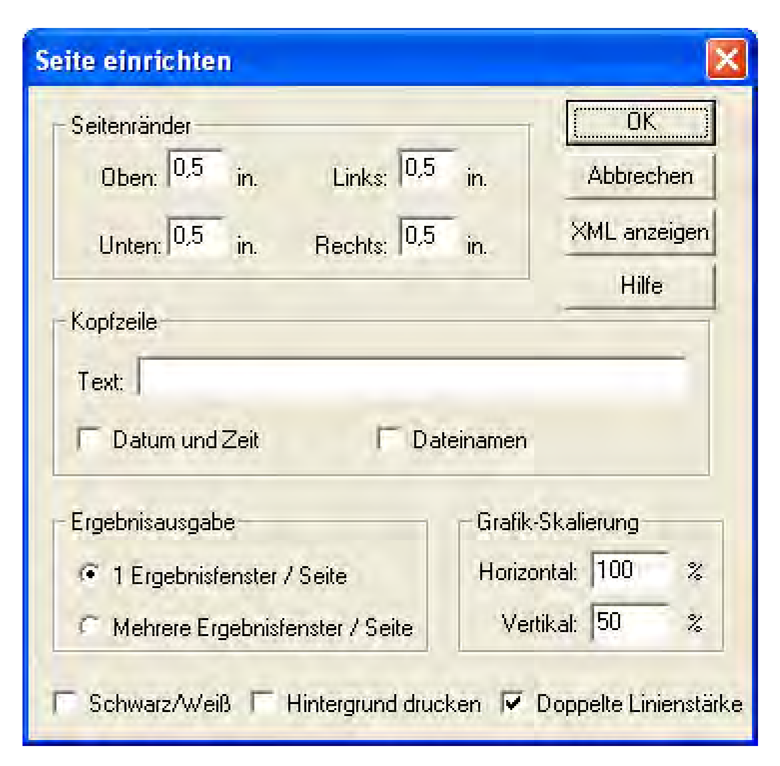


Abbildung 3.16: Dialogbox zum Seite einrichten

In dieser Dialogbox können Sie:

1. die *Druckränder* der Seite festlegen.

2. eine *Kopfzeile* angeben, die auf jeder Seite gedruckt wird.

3. angeben, ob jedes Ergebnisfenster (Text oder Grafik) auf einer separaten Seite gedruckt werden soll, oder ob, wenn möglich, *mehrere* Ergebnisfenster auf einer Seite gedruckt werden.

4. die *Größe der Diagramme* im Verhältnis zur Größe der Seite bestimmen (in %).

5. angeben, ob die Ergebnisse in *Schwarz/Weiß* gedruckt werden, auch wenn Sie einen Farbdrucker verwenden.

6. angeben, ob ein farbiger *Hintergrund* eines Diagramms, soweit vorhanden, mit gedruckt wird.

7. bestimmen, ob die *Linien in doppelter Linienstärke* – 2 Pixel breit anstatt 1 Pixel – gedruckt werden. Ein hoch aufgelöster Druck kann in dieser Option zu fett erscheinen.

Weitere Optionen, wie z. B. Hoch- oder Querformat, werden im *Datei*-Menü unter *Druckeinrichtung* eingestellt, wodurch auf die zu Ihrem Druckertreiber gehörige Dialogbox zugegriffen wird.

3.4 Publizieren der Ergebnisse

Die Ergebnisse einer statistischen Analyse können in Form von HTML-Dokumenten exportiert werden, um sie in einem Webbrowser anzuzeigen. Hierfür wählen Sie *StatPublish* im *Datei*-Menü. Über den Webbrowser können dann auch Mitarbeiter und andere Personen, die STATGRAPHICS Centurion XVI nicht auf ihrem Computer installiert haben, Ihre Analysen verfolgen. Die Optionen des Publizierens werden in Kapitel 5 beschrieben.

Sie können die Analysen auch in den StatReporter kopieren, um sie zu kommentieren und als RTF (RichText Format) zu sichern. RTF-Dokumente können von anderen Anwendungen wie Microsoft Word gelesen werden. Die Verwendung des StatReporters wird in Kapitel 6 beschrieben.

Kapitel

4

Diagramme

Bearbeiten der Diagramme, Sichern der Grafik-Profile, Interaktives Arbeiten mit Diagrammen, Sichern der Diagramme in Bilddateien und Kopieren der Diagramme in andere Anwendungen.

Mit den mehr als 160 statistischen Prozeduren in STATGRAPHICS Centurion XVI können Hunderte unterschiedlicher Diagrammtypen erstellt werden. Um den Prozess der Datenanalyse zu vereinfachen, werden den Diagrammen Standards wie Titel, Skalierung und ähnliches zugewiesen. Für die Analyse selbst genügen diese Standardangaben in den meisten Fällen. Eine Publikation der Ergebnisse erfordert hingegen Diagramme, deren Erscheinungsbild höchsten Ansprüchen gerecht wird.

Dieses Kapitel beschreibt, wie Diagramme in STATGRAPHICS Centurion XVI bearbeitet werden können. Es zeigt, wie man sie für eine Publikation vorbereitet, wie man sie in Anwendungen (z. B. Microsoft Word und PowerPoint) einfügt und wie mit ihnen interaktiv gearbeitet werden kann. Zum Beispiel, wenn Sie einen interessanten Datenpunkt in Ihrem Diagramm finden und mehr über ihn wissen möchten. Oder wenn Sie ein 3D-Diagramm rotieren möchten, um ein Gefühl für die Beziehungen zwischen den Variablen der X-, Y- und Z-Achsen zu bekommen.

Als Beispiel werden wir erneut die Daten der *93cars.sgd*-Datei verwenden. Zum Einstieg sollen mit dem Diagramm des angepassten Modells, das die Reichweite *MPG City* und das Fahrzeuggewicht *Weight* in Beziehung zueinander setzt, einige der wichtigsten Grafik-Bearbeitungsoptionen veranschaulicht werden.

4.1 Bearbeiten der Diagramme

Um Kurven anzupassen, die eine abhängige Variable Y (Zielgröße) und eine unabhängige Variable X (erklärende oder Einflussvariable) in Beziehung setzen, wird gewöhnlich die *Einfache Regression* angewendet. Wie im vorherigen Kapitel beschrieben, ist das S-Kurven-Modell besonders geeignet, um die Beziehung zwischen *MPG City* und *Weight* der *93cars.sgd*-Datei darzustellen.

Das Diagramm des angepassten S-Kurven-Modells sieht, wenn es erstmalig erstellt wird, wie folgt aus:

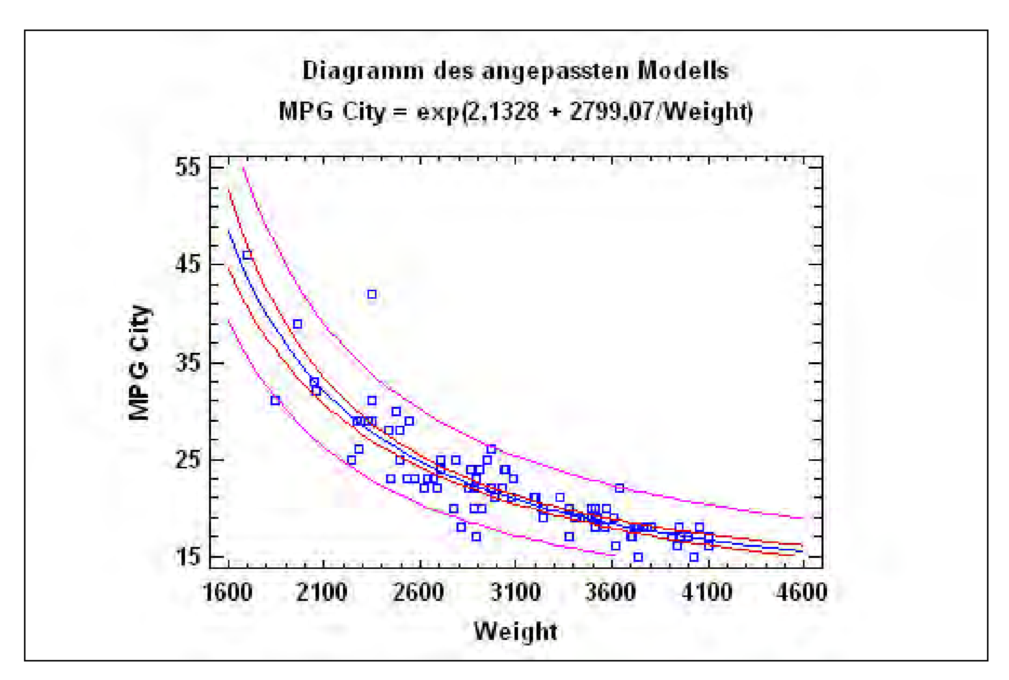


Abbildung 4.1: Diagramm des angepassten Modells mit Standardüberschrift und -skalierung

Überschriften, Skalierung, Punkt- und Linienarten, Farben und weitere Eigenschaften wurden automatisch erzeugt.

4.1.1 Layout

Um das erzeugte Diagramm nach Ihren Vorstellungen zu verändern, maximieren Sie zunächst das Ergebnisfenster. Wählen Sie dann die Schaltfläche *Diagramm-Optionen* auf der Analyse-Symbolleiste aus. Dies öffnet eine Dialogbox mit Registern, die sich auf jeweils ein Diagramm-Element beziehen.

Über das Register *Layout* in der *Diagramm-Optionen*-Dialogbox können einige Grundeigenschaften des Diagramms verändert werden:

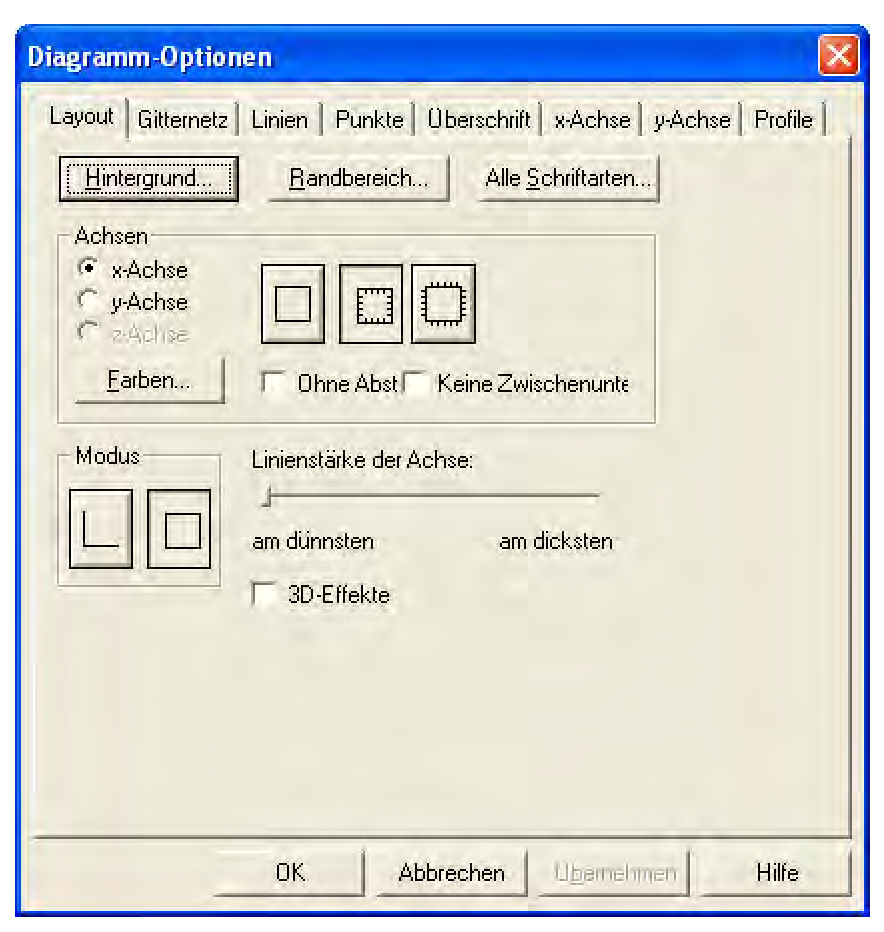


Abbildung 4.2: Layout-Register in der Dialogbox Diagramm-Optionen

Dies schließt z. B. die Ausrichtung der *Achseneinteilung*, die Linienstärke der Achsen sowie die Hintergrundfarbe und den Rahmen des Diagramms ein. Wenn Sie z.B. die Farbe des *Hintergrunds*

gelb einstellen, das Kontrollkästchen *3D-Effekte* markieren und die Einstellungen mit *Übernehmen* bestätigen, wird das Diagramm folgendermaßen verändert:

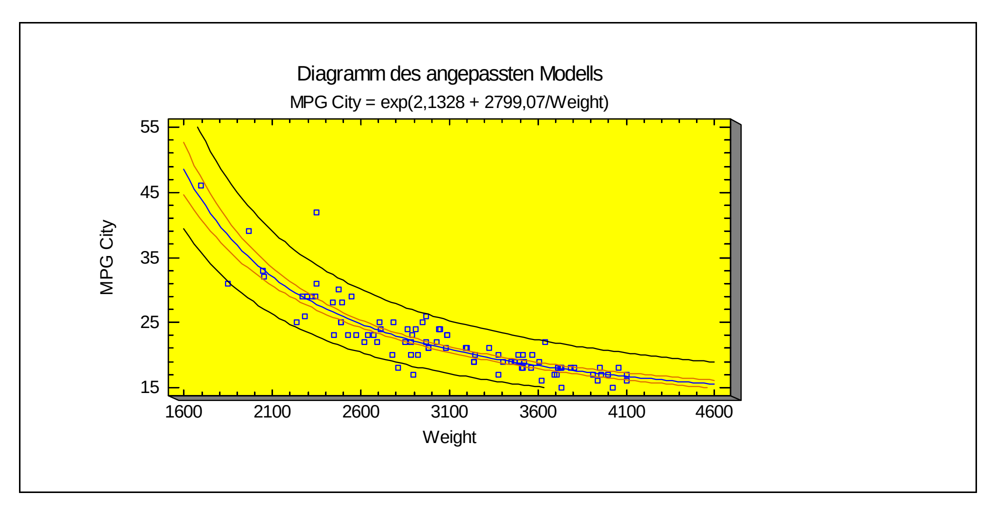


Abbildung 4.3: Diagramm nach dem Einstellen der Hintergrundfarbe und der 3D-Effekte

> ANMERKUNG: Die Veränderungen der Farbgebung können Sie in der Hilfe-Dokumentation nachschlagen, die in Ihrem Programm z. B. unter *Hilfe – Online Manuals* enthalten ist.

4.1.2 Gitternetz

Das Register *Gitternetz* wird verwendet, um dem Diagramm ein Raster aus Gitternetzlinien hinzuzufügen:

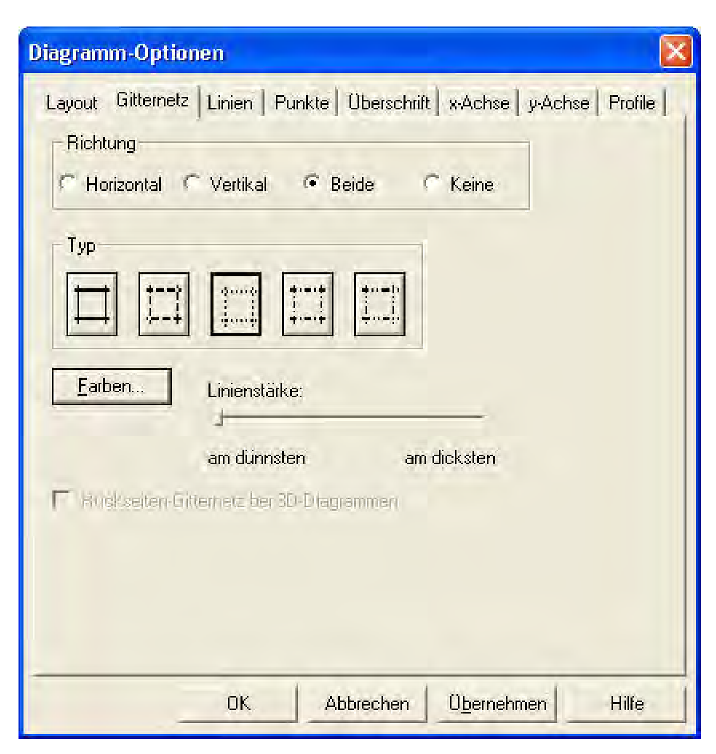


Abbildung 4.4: Gitternetz-Register in der Dialogbox Diagramm-Optionen

Wenn Sie hier ein graues gestricheltes Gitternetz in der Richtung „*Beide*“ wählen, sieht das Diagramm folgendermaßen aus:

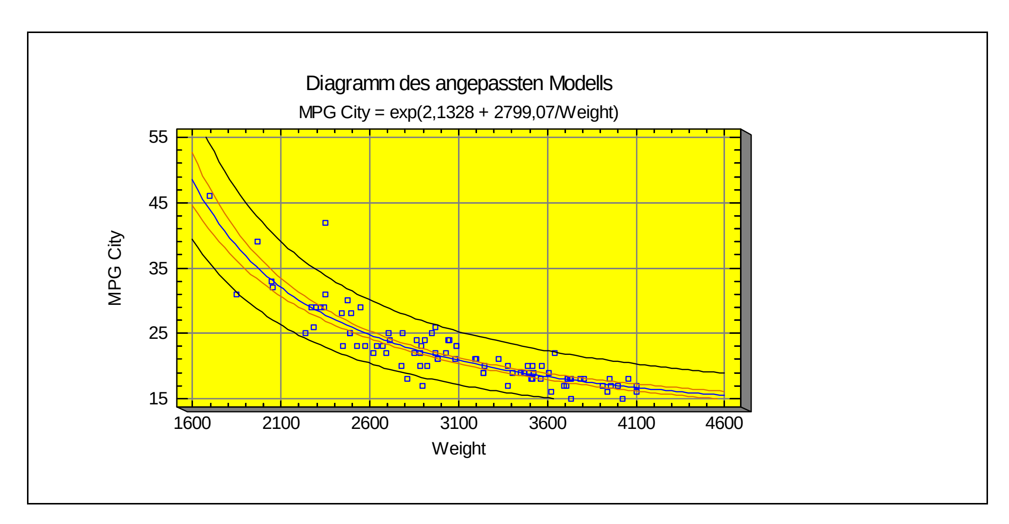


Abbildung 4.5: Diagramm nach dem Einfügen eines Gitternetzes

4.1.3 Linien

Mit dem Register *Linien* können die Art der Linien eines Diagramms, Linienfarbe und –stärke festgelegt werden:

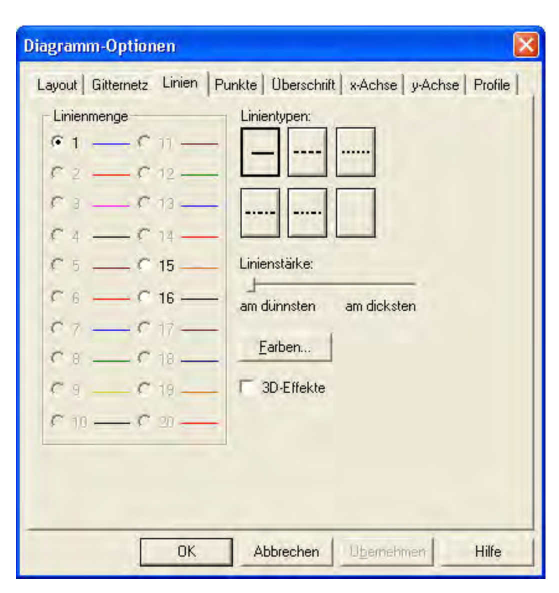


Abbildung 4.6: Linien-Register in der Dialogbox Diagramm-Optionen

Für ein Diagramm wie das des angepassten Modells gibt es drei Liniengruppen: die mittlere Linie (Regressionskurve), die inneren Konfidenzgrenzen und die äußeren Vorhersagegrenzen. Um eine, zwei oder alle dieser Linien zu ändern, markieren Sie die entsprechenden Optionsfelder *1*, *2* und/oder *3* in dem Gruppenfeld *Linienart* und legen dann die gewünschten Eigenschaften fest. Wenn Sie die *Linienstärke* der Mittellinie erhöhen und die *Typen* der anderen Linien verändern, erhalten Sie folgendes Diagramm:

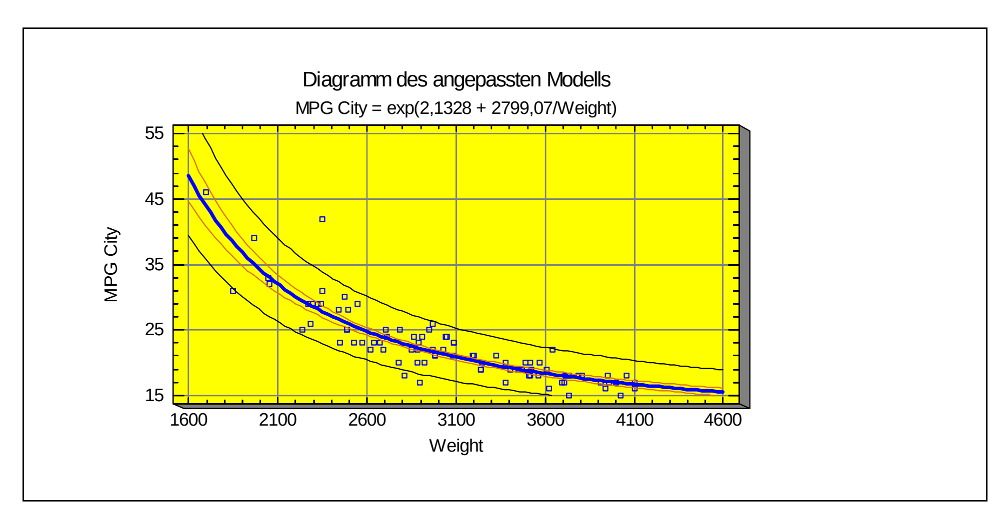


Abbildung 4.7: Diagramm nach dem Bearbeiten der Linientypen

> ANMERKUNG: Die Linienstärke kann nur bei durchgängigen Linientypen verändert werden.

4.1.4 Punkte

Mit dem Register *Punkte* können die Art der Datenpunkte eines Diagramms sowie ihre Farbe und Größe bestimmt werden:

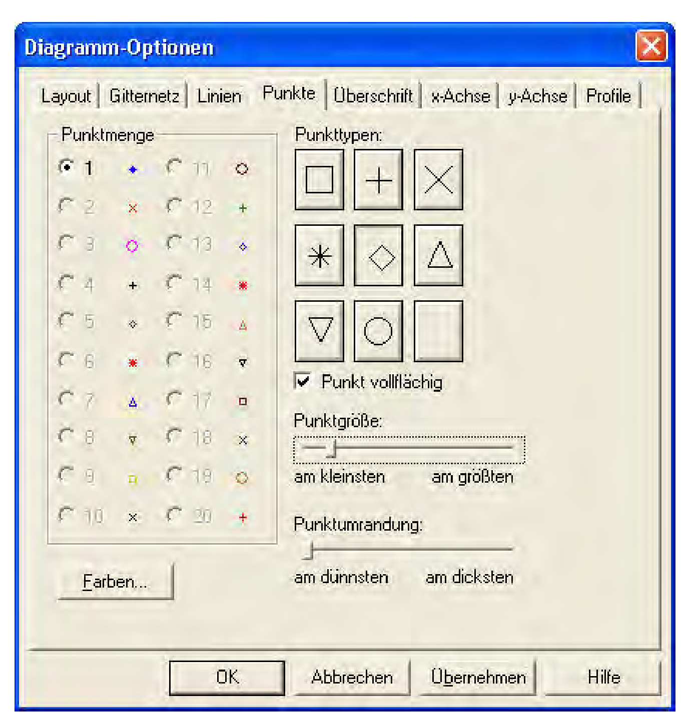


Abbildung 4.8: Punkte-Register in der Dialogbox Diagramm-Optionen

Das Optionsfeld *1* in dem Gruppenfeld *Punktmenge* weist der ersten Punktart im Diagramm die gewählten Eigenschaften zu. In unserem Beispiel gibt es nur eine Art von Datenpunkten. Wenn Sie als *Punkttyp* farbig gefüllte Rauten (*vollflächig*) wählen, erhalten Sie folgendes Diagramm:

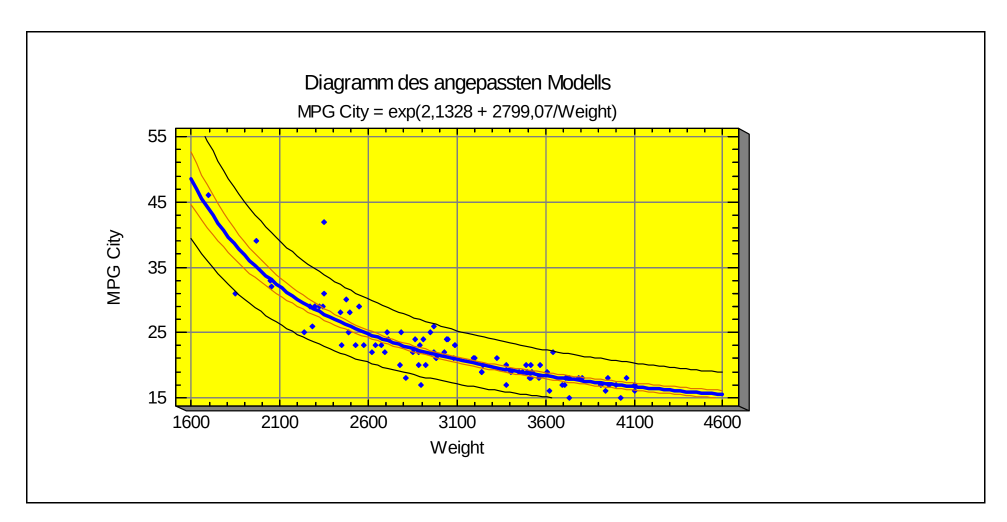


Abbildung 4.9: Diagramm nach dem Bearbeiten der Punkttypen

4.1.5 Diagramm-Überschrift

Mit dem Register *Überschrift* können der Inhalt und die Schriftart des Titels, der über dem Diagramm angezeigt wird, eingestellt werden:

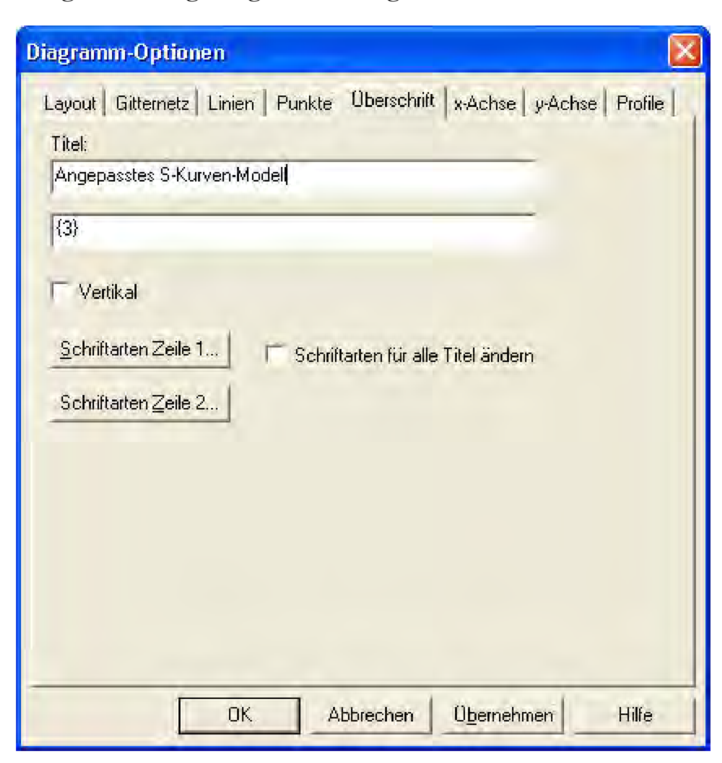


Abbildung 4.10: Überschrift-Register in der Dialogbox Diagramm-Optionen

Der Titel des Diagramms kann bis zu zwei Zeilen umfassen. Ein Eintrag wie "{3}" in einem *Titel*-Eingabefeld zeigt an, dass der Text automatisch während der Analyse generiert wurde und in der Regel die Variablennamen oder die errechneten Statistiken enthält. Sie können jeden Titel, einschließlich der automatisch generierten, verändern. Zudem können Sie den Titel mit der Maus an eine gewünschte Stelle des Diagramms bewegen:

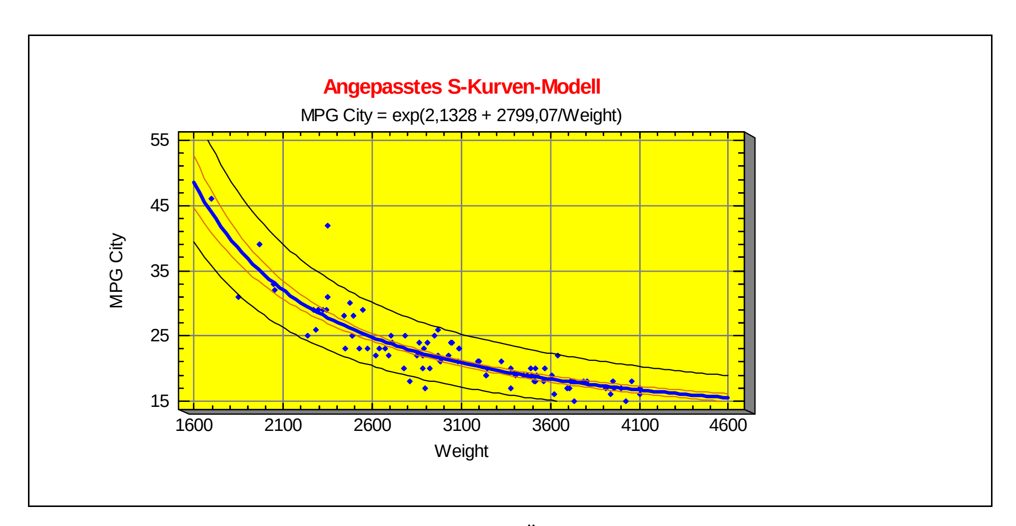


Abbildung 4.11: Diagramm nach dem Bearbeiten der Überschrift

4.1.6 Achsenskalierung

Die Dialogbox *Diagramm-Optionen* enthält auch Register (hier: *x-Achse* und *y-Achse)*, über die man die Bezeichnungen der Diagrammachsen und ihre Skalierung einstellen kann:

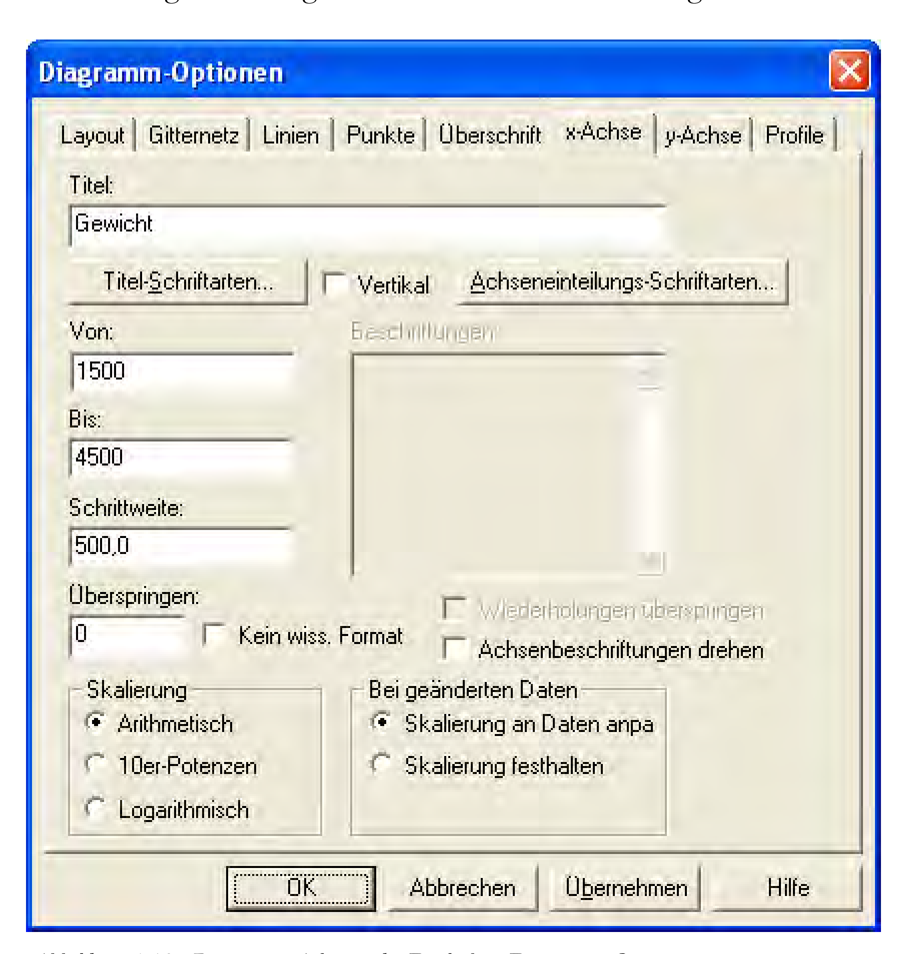


Abbildung 4.12: Register x-Achse in der Dialogbox Diagramm-Optionen

Wichtige Eingabefelder auf diesen Registern sind:

1. *Titel:* Bezeichnung, die entlang der Achse angezeigt wird.

2. *Von*, *Bis*, *Schrittweite* und *Überspringen:* stellt die Achseneinteilung ein. Der Wert in *Überspringen* wird verwendet, um zu verhindern, dass sich die Einheiten der Achse überlappen. Wenn hier der Wert *1* angegeben ist, wird jede zweite Einheit nicht angezeigt.

3. *Achsen-Beschriftungen drehen:* richtet die Einteilungseinheiten der Achse vertikal aus.

4. *Kein wiss. Format:* gibt sehr große und sehr kleine Zahlen im Standardformat und nicht im wissenschaftlichen Format an.

5. *Skalierung:* zeichnet die Achse mit zwei verschiedenen 10er-Logarithmus-Einteilungen.

6. *Bei geänderten Daten:* gibt an, ob die Skalierung erhalten bleibt oder bei der Ausgabe neuer Daten angepasst wird.

7. *Schriftarten:* hier können die Farbe, Größe oder Stil der Beschriftungen und Achsenunterteilungen eingestellt werden.

Mit den Angaben in der obigen Dialogbox wurde das folgende Diagramm erstellt:

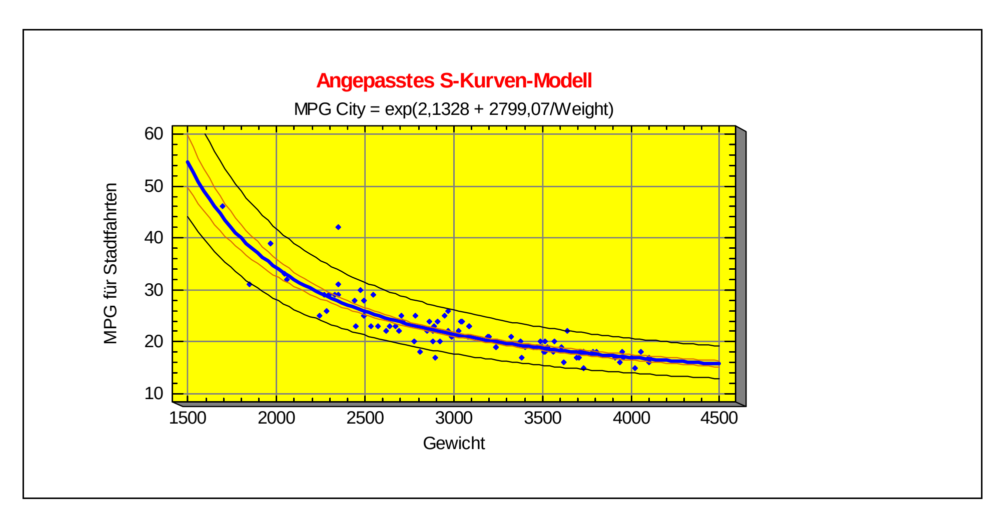


Abbildung 4.13: Diagramm nach dem Bearbeiten der Achsenbezeichnungen und -einteilungen

4.1.7 Füllmuster

Einige Diagramme, wie z. B. Häufigkeitsdiagramme, enthalten flächige Bereiche. Mit dem Register *Füllmuster* in der *Diagramm-Optionen*-Dialogbox können die Farbe und der Fülleffekt für die Flächen, wie z.B. Säulen/Balken, Polygone und Kreissegmente, festgelegt werden.

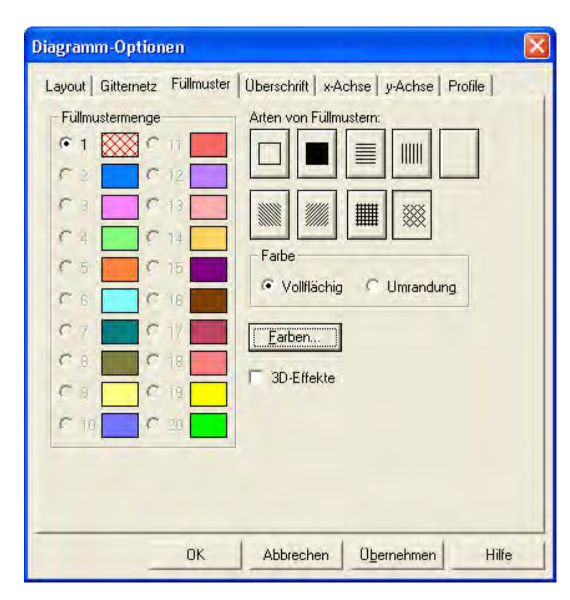


Abbildung 4.14: Register Füllmuster in der Dialogbox Diagramm-Optionen

Wenn das Optionsfeld *1* im Gruppenfeld *Füllmuster* markiert ist, werden nur die Fülleffekte der ersten Flächenart im Diagramm bearbeitet. In einem Häufigkeitsdiagramm sind alle Flächen der ersten Flächenart zugeordnet. In Diagrammtypen wie dem Kreisdiagramm können mehrere unterschiedliche Fülleffekte eingestellt werden. In diesen Fällen steuern die Optionsfelder *2* bis *20* die weiteren Fülleffekte.

Wenn Diagrammtypen wie das Häufigkeitsdiagramm in Schwarz/Weiß gedruckt werden, kann es günstig sein, die Flächen mit Schraffuren zu füllen:

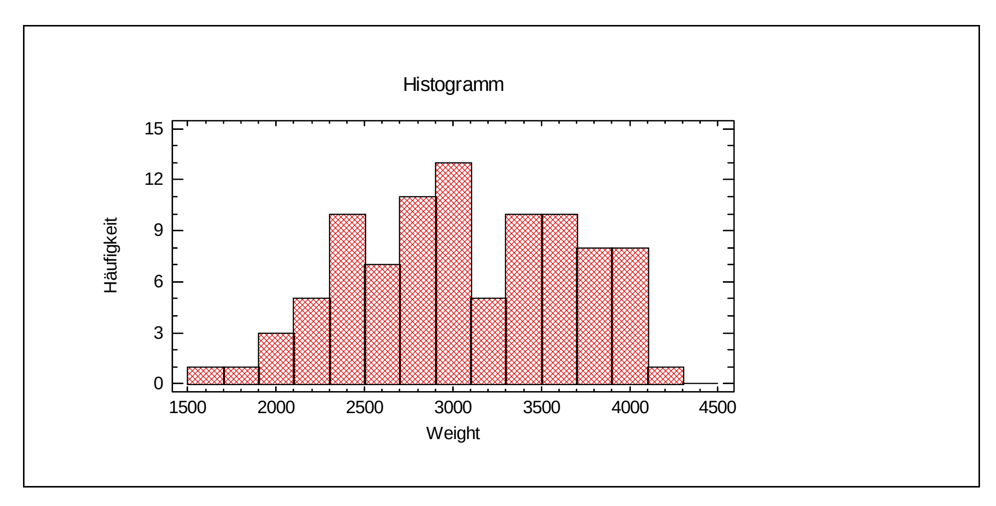


Abbildung 4.15: Häufigkeitsdiagramm mit bearbeitetem Fülleffekt

4.1.8 Text, Beschriftungen und Legenden

Für Diagramme, die zusätzliche Beschriftungen oder Legenden enthalten, werden in der Dialogbox *Diagramm-Optionen* weitere Register angezeigt, über die man deren Text und Schriftarten verändern kann.

4.1.9 Hinzufügen neuer Texte

Jedem Diagramm können über die Schaltfläche *Text hinzufügen* auf der Analyse-Symbolleiste weitere Textfelder zugefügt werden. Hinter der Schaltfläche verbirgt sich eine Dialogbox, mit der ein Text eingegeben werden kann:

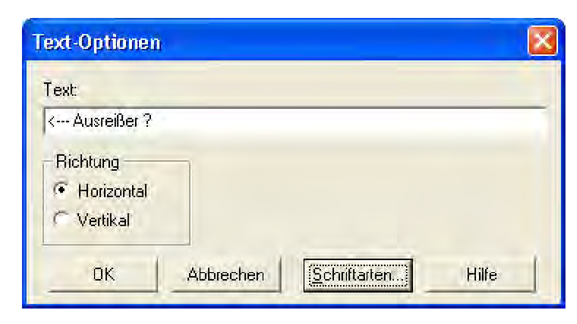


Abbildung 4.16: Dialogbox zum Einfügen neuer Texte oder Beschriftungen

Der eingegebene Text wird zunächst automatisch unter dem Titel des Diagramms positioniert. Mit der Maus kann er jedoch an jede beliebige Stelle in der Grafik bewegt werden:

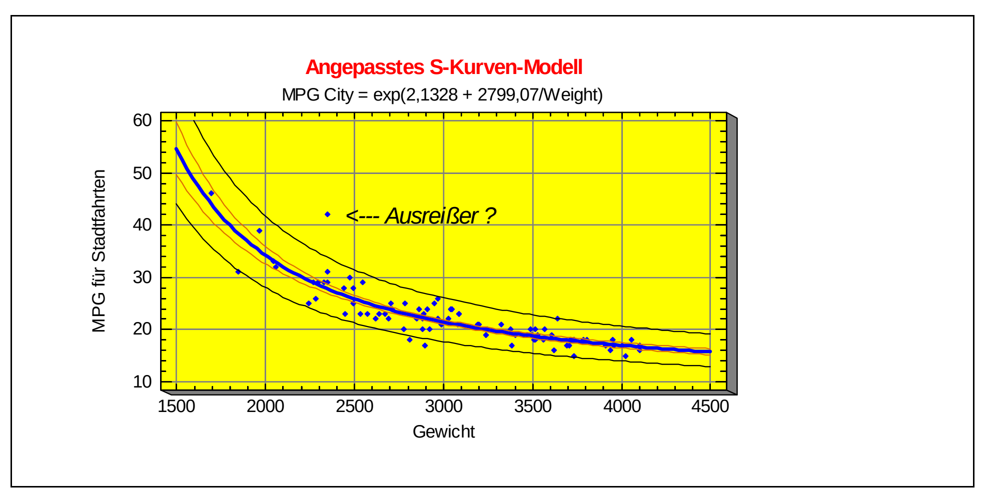


Abbildung 4.17: Diagramm nach dem Einfügen eines neuen Texts

Wenn Sie nach dem Einfügen des Texts weitere Veränderungen daran vornehmen möchten, markieren Sie den Text und wählen die Schaltfläche *Diagramm-Optionen*.

4.2 Auseinanderziehen eines Streudiagramms

Wenn ein Streudiagramm mit einer oder zwei diskreten Variablen erstellt wird, können sich einige der Datenpunkte an der gleichen Stelle befinden und demnach einige der Werte verdeckt sein. Auf der Analyse-Symbolleiste befindet sich die Schaltfläche *Auseinanderziehen*, die dieses Problem löst, indem die überlappenden Punkte in vertikaler und/oder horizontaler Richtung versetzt werden. Als Beispiel hierfür betrachten Sie das folgende Diagramm aus den Daten der *93cars.sgd*-Datei:

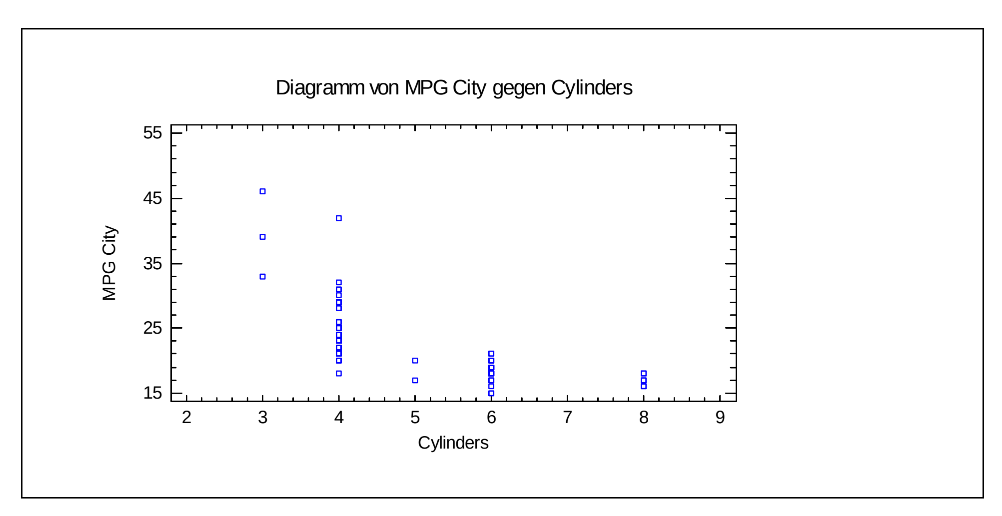


Abbildung 4.18: Streudiagramm aus MPG City abgebildet auf Cylinders

Obwohl die Datei 93 Zeilen besitzt, zeigt das Diagramm weitaus weniger Punkte.

Wenn Sie nun die Schaltfläche *Auseinanderziehen* wählen, öffnet sich eine Dialogbox, über die Sie die Datenpunkte zufällig auseinanderziehen können:

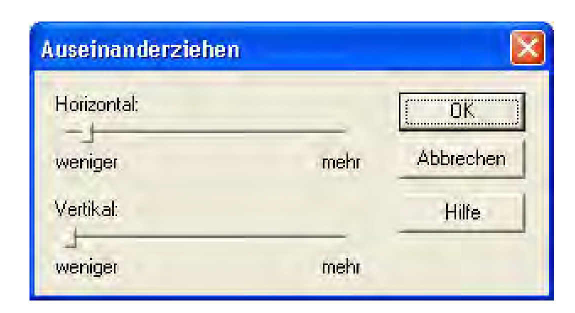


Abbildung 4.19: Auseinanderziehen-Dialogbox

Schon wenn Sie die Punkte in horizontaler Richtung leicht versetzen, ist die Verteilung der Punkte viel besser nachzuvollziehen:

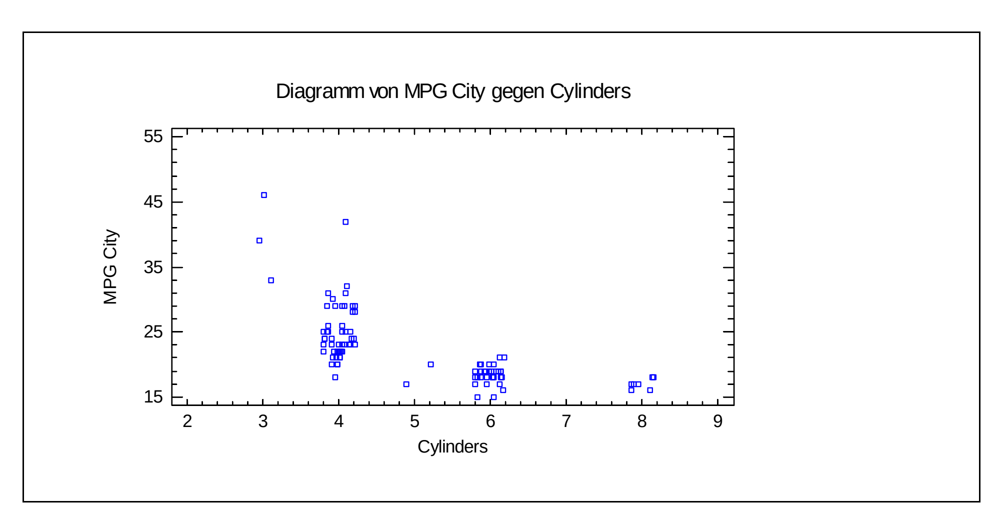


Abbildung 4.20: Streudiagramm nach dem Auseinanderziehen der Punkte in horizontaler Richtung

Jeder der Punkte wurde in einem zufälligen Abstand zur ursprünglichen Position entlang der horizontalen Achse versetzt. Das *Auseinanderziehen* der Punkte betrifft ausschließlich ihre Darstellung im Diagramm, die Daten im Datenblatt oder Berechnungen aus ihnen werden dadurch nicht verändert.

4.3 Einfärben der Punkte eines Streudiagramms

Die Beziehungen zwischen den Variablen können häufig besser visualisiert werden, indem die Datenpunkte eines Streudiagramms den Werten einer weiteren Variable entsprechend eingefärbt werden. Betrachten Sie als Beispiel das folgende Matrix-Diagramm für einige ausgewählte Variablen der *93cars.sgd*-Datei:

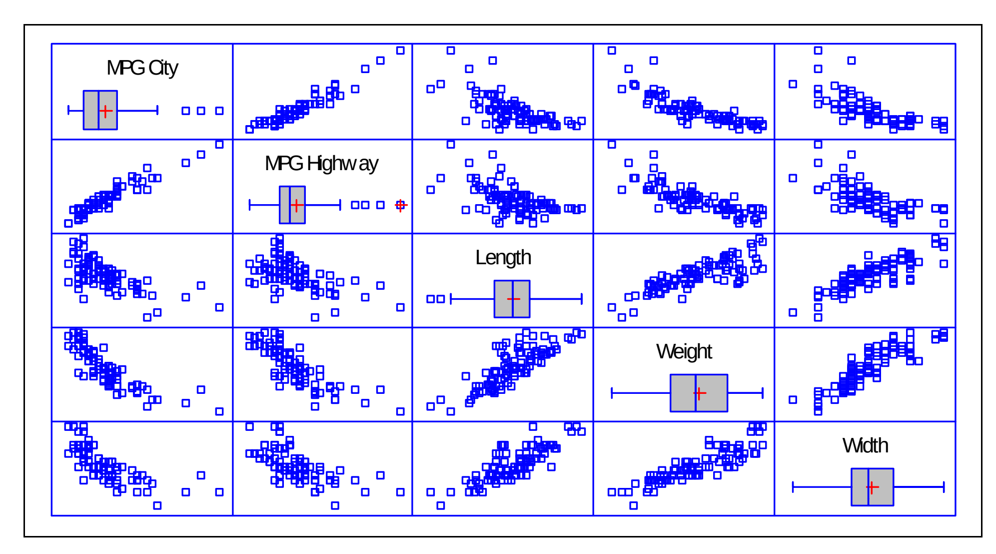


Abbildung 4.21: Matrix-Diagramm für die Daten aus der 93cars.sgd-Datei

Die Streudiagramme in jeder Zelle der Matrix bilden die Werte der Variablen ihren Zeilen und Spalten entsprechend ab.

Angenommen, Sie möchten zeigen, in welcher Beziehung die PS-Leistung (*Horsepower*) der Fahrzeuge zu den fünf abgebildeten Variablen steht. Wenn Sie die Schaltfläche *Einfärben* auf der Analyse-Symbolleiste auswählen, wird die folgende Dialogbox aufgerufen:



Abbildung 4.22: Dialogbox zum Auswählen einer Variable für das Einfärben

Wählen Sie eine quantitative Variable, um die Punkte zu kodieren. Nachdem Sie eine Variable für das Einfärben ausgewählt haben, erscheint folgende Dialogbox:



Abbildung 4.23: Dialogbox zum Auswählen des Einfärbe-Intervalls

Mit den beiden Schiebebalken werden die untere und obere Grenze für die Variable eingestellt. Es werden alle Punkte des Diagramms rot eingefärbt, die im Bereich des hier angegebenen Intervalls liegen. In dem folgenden Beispiel-Diagramm wurden alle Punkte, welche die Fahrzeuge mit einer PS-Leistung zwischen 55,0 und 121,5 repräsentieren, rot dargestellt:

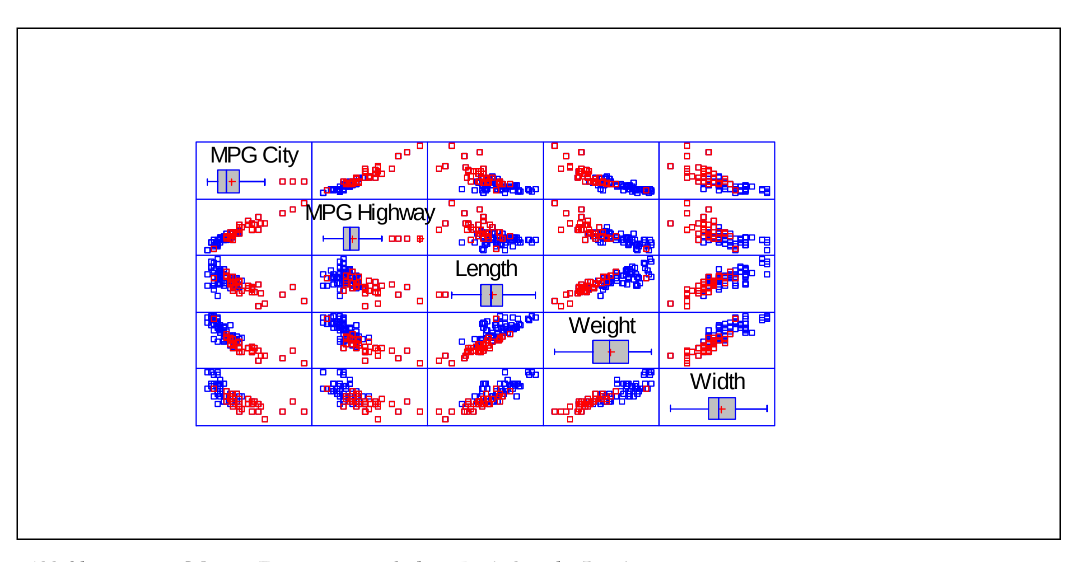


Abbildung 4.24: Matrix-Diagramm nach dem Einfärben der Punkte

Aus dieser Darstellung wird deutlich, dass die PS-Leistung (*Horsepower)* stark mit den anderen Variablen korreliert.

4.4 Glätten eines Streudiagramms

Um die Beziehungen zwischen den Variablen eines Streudiagramms deutlicher abzubilden, kann eine Glättung verwendet werden. Wählen Sie hierfür die Schaltfläche *Glätten/Rotieren* auf der Analyse-Symbolleiste aus. Dies ruft die folgende Dialogbox auf:



Abbildung 4.25: Dialogbox zum Glätten eines Streudiagramms

Um ein Streudiagramm zu glätten, wird eine Reihe von Positionen auf der x-Achse ausgewählt. An jeder der Positionen wird das gewichtete Mittel des vorgegebenen, zu glättenden Anteils an Punkten gezeichnet, die am dichtesten um die gewählten Stellen liegen. Zu den besten Glättungsmethoden gehört LOWESS (LOcally WEighted Scatterplot Smoothing), gewöhnlich mit einem Glättungsanteil zwischen 40 % und 60 %. Das Ergebnis der Glättung des *Matrix-Diagramms* für die Fahrzeug-Daten stellt sich wie folgt dar:

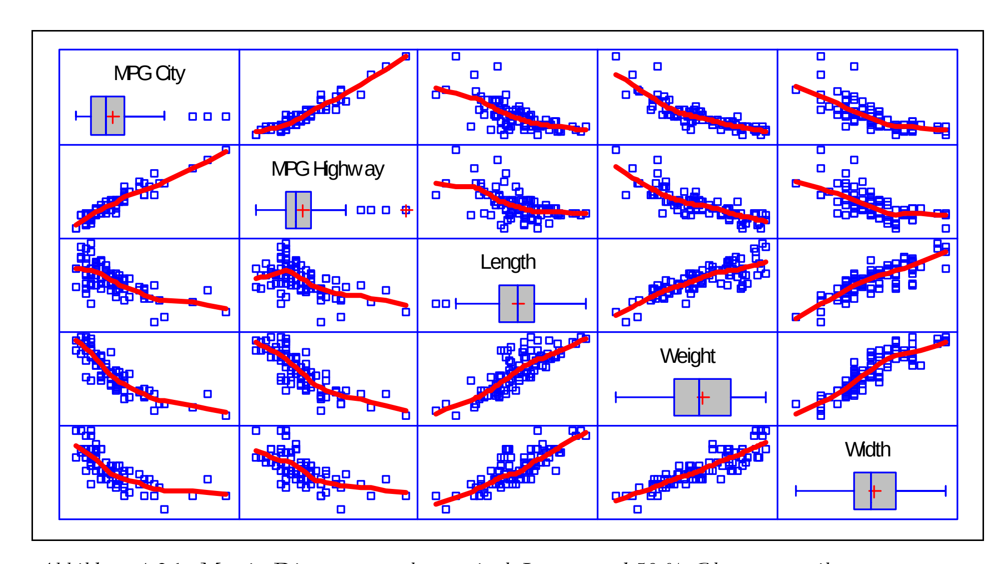


Abbildung 4.26: Matrix-Diagramm, geglättet mittels Lowess und 50 % Glättungsanteil

Die Glättung hilft dabei, die Art der Beziehungen zwischen den Variablen zu illustrieren.

4.5 Identifizieren der Punkte eines Diagramms

Um die Zeilennummer und die Koordinaten eines beliebigen Datenpunkts in einem Diagramm anzuzeigen, wählen Sie den entsprechenden Punkt mit der linken Maustaste an. Bei gedrückter Maustaste öffnet sich ein kleines Feld in der rechten oberen Ecke des Diagramms, das die Zeilennummer und die Koordinaten *X* und *Y* des Punkts angibt:

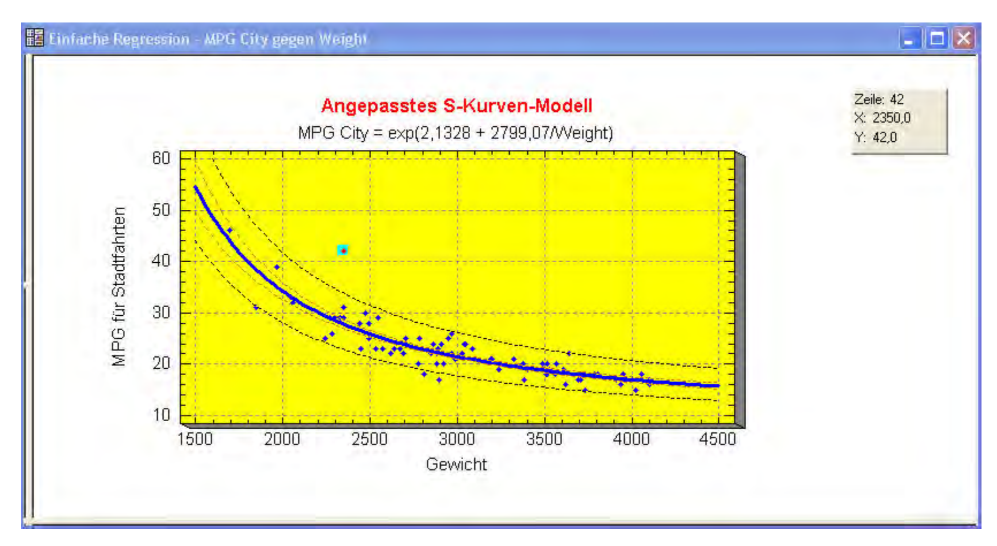


Abbildung 4.27: Anzeige der Informationen zu einem markierten Punkt des Diagramms

Gleichzeitig wird die Zeilennummer auch im Feld *Zeile* auf der Analyse-Symbolleiste angezeigt:

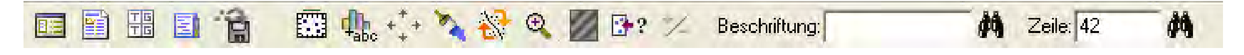


Abbildung 4.28: Anzeige der Zeilennummer eines markierten Punkts auf der Analyse-Symbolleiste

Weitere Informationen zu dem markierten Datenpunkt kann man erhalten, indem man die Schaltfläche *Identifizieren* auf der Analyse-Symbolleiste drückt und in der aufgerufenen Dialogbox einen Spaltennamen des Datenbuchs in das Eingabefeld *Identifizieren durch* einfügt:

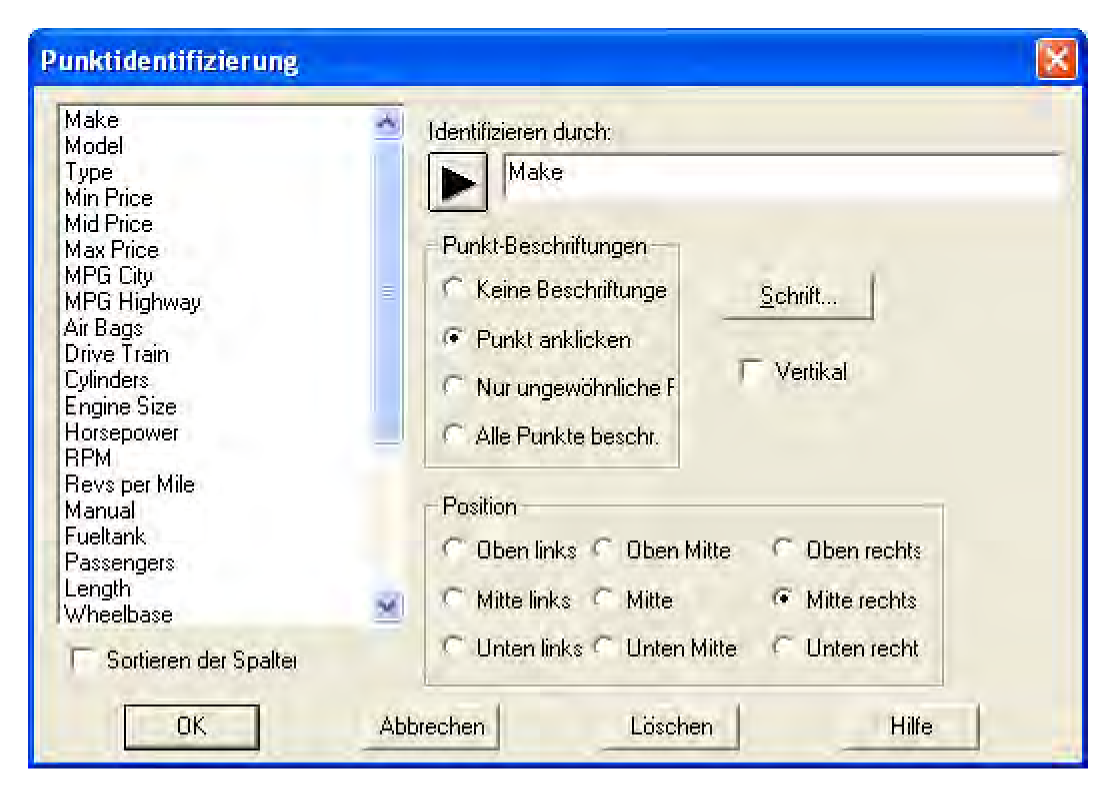


Abbildung 4.29: Dialogbox zum Identifizieren eines Punkts aus dem Diagramm

Nach Auswahl der Variable wird jedes Mal, wenn Sie einen Punkt im Diagramm markieren, ein Wert im Feld *Beschriftung* auf der Analyse-Symbolleiste angezeigt:

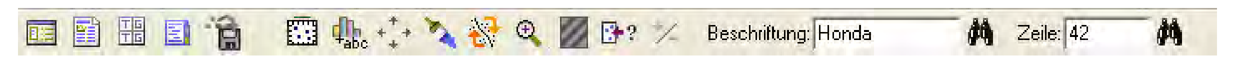


Abbildung 4.30: Analyse-Symbolleiste mit der Anzeige von „Make" zum markierten Punkt

Die Fernglas-Schaltflächen rechts neben den Eingabefeldern *Beschriftung* und *Zeile* dienen dazu, bestimmte Punkte in einem Diagramm zu lokalisieren. Wenn Sie einen Wert in eines der Felder eingeben und die zugehörige *Lokalisieren*-Schaltfläche anklicken, werden alle Punkte, die dem eingegebenen Wert entsprechen, im Diagramm markiert. Das folgende Diagramm zeigt die Punkte, die für alle Hondas stehen, rot markiert an:

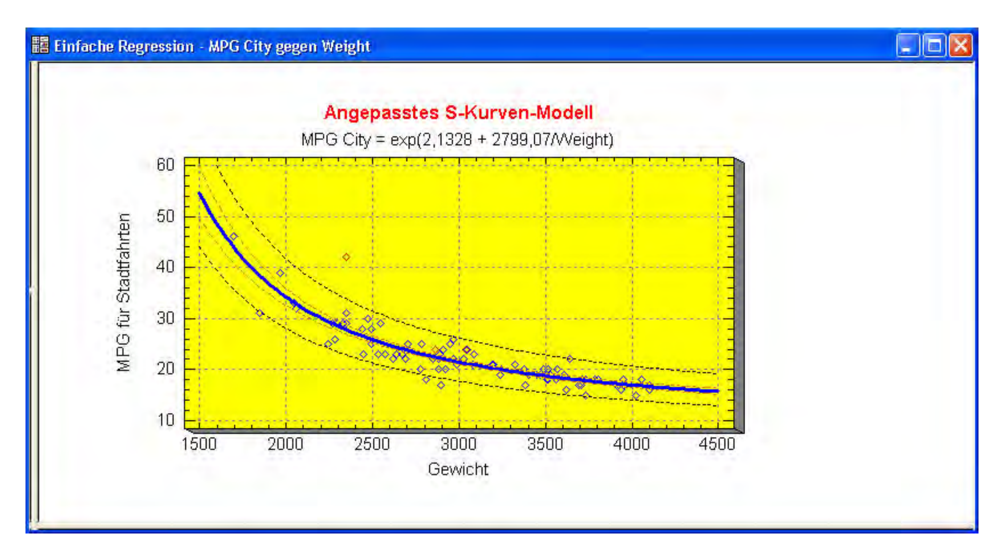


Abbildung 4.31: Diagramm mit den markierten Punkten aller Hondas

Diese Technik hat sich auch für *Matrix-Diagramme* bewährt. Im folgenden Bild sind alle Punkte, die Zeile 42 des Datenblatts entsprechen, rot hervorgehoben:

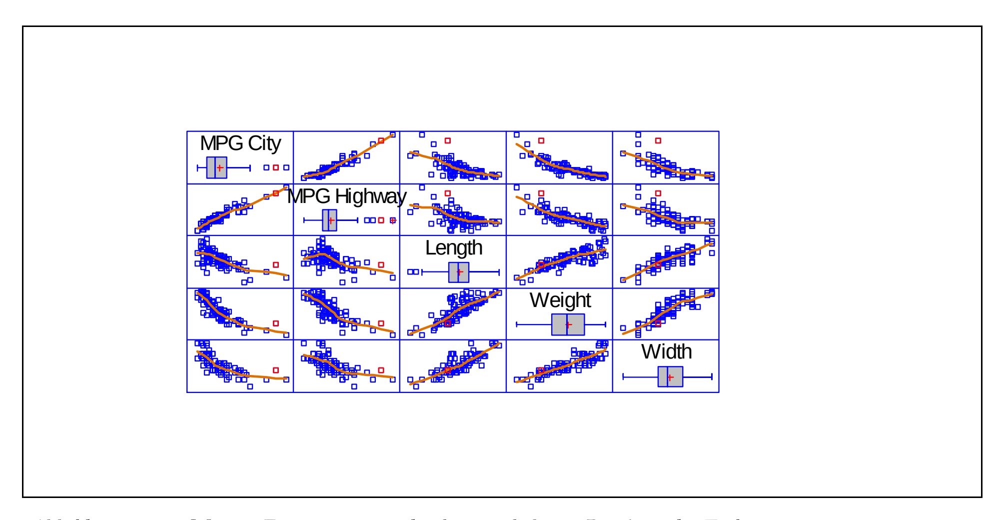


Abbildung 4.32: Matrix-Diagramm mit den hervorgehobenen Punkten der Zeile 42

Durch das Lokalisieren eines Punkts in einem Matrix-Diagramm, das mehrere Variablen gegenüberstellt, kann man besser einschätzen, ob der Punkt einen Ausreißer darstellt.

ANMERKUNG: Die Farbe zum Hervorheben der Punkte kann im Register *Diagramme* der Dialogbox *Voreinstellungen,* die man im *Bearbeiten*-Menü findet, festgelegt werden.

4.6 Einfügen der Diagramme in andere Anwendungen

Sobald ein Diagramm in STATGRAPHICS Centurion XVI erstellt wurde, kann es auf einfache Weise in andere Programme wie Microsoft Word oder PowerPoint kopiert werden:

1. Maximieren Sie das Ergebnisfenster mit dem Diagramm.

2. Wählen Sie *Kopieren* aus dem *Bearbeiten*-Menü von STATGRAPHICS Centurion XVI.

3. Wählen Sie in der anderen Anwendung: *Einfügen*.

Die Diagramme werden standardmäßig im Bild-Format eingefügt, das einer Windows-Metadatei entspricht. Wenn Sie ein Diagramm in einem anderen Format einfügen möchten, wählen Sie *Speziell einfügen* anstelle von *Einfügen*.

Um eine vollständige Analyse mit allen Tabellen und Diagrammen in eine andere Anwendung einzufügen, kopieren Sie die Analyse über das Kontextmenü des Analyse-Fensters zunächst in den StatReporter. Anschließend kopieren Sie den StatReporter in die andere Anwendung. Der genaue Vorgang wird in Kapitel 7 beschrieben.

Um ein Diagramm einschließlich des umgebenden Fensters zu kopieren, wird ein Programm für die Erstellung von Screenshots benötigt. Bei der Herstellung dieses Handbuches wurde das Programm *SnagIt* verwendet, das man unter *www.techsmith.com* kaufen kann. Falls Sie *SnagIt* benutzen, empfehlen wir, die *Input*-Option auf "Fenster" und die *Output*-Option auf "Zwischenablage" einzustellen. Dann können Sie die Bilder direkt in ein beliebiges Dokument einfügen.

4.7 Speichern der Diagramme als Bilddateien

Einzelne Diagramme können in Form einer Bilddatei gesichert werden, indem Sie das Diagramm maximieren und im *Datei*-Menü *Grafik speichern* auswählen. In der aufgerufenen Dialogbox können Sie den Dateinamen und das Bildformat festlegen:

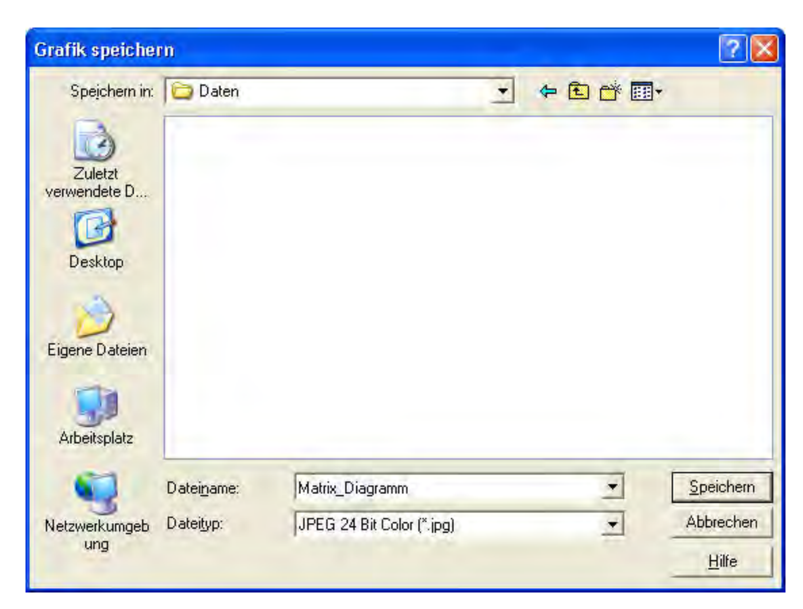


Abbildung 4.33: Dialogbox zum Speichern eines Diagramms als Bilddatei

Zum Sichern von Diagrammen, die in Word oder PowerPoint eingefügt werden sollen, bietet das Windows-Metafiles-Format größte Flexibilität. Soll das Diagramm auf einer Internetseite abgebildet werden, wird empfohlen, es als JPEG-Datei zu speichern.

Kapitel

5

StatFolios

Sichern Ihrer Sitzung, Publizieren der Ergebnisse im HTML-Format und Automatisieren von Analysen mit Hilfe von Start-Scripts.

Jedes Mal, wenn Sie eine statistische Analyse aus dem STATGRAPHICS Centurion XVI-Menü auswählen, wird ein neues Analyse-Fenster geöffnet. Sie können alle Analyse-Fenster auf einmal sichern, indem Sie ein *StatFolio* anlegen. Ein StatFolio ist eine Datei, welche die Definitionen aller erstellten statistischen Analysen sowie alle benötigten Verweise auf die dafür verwendeten Daten enthält. Indem Sie ein StatFolio speichern und es später wieder öffnen, können Sie Ihre aktuelle STATGRAPHICS Centurion XVI-Sitzung effektiv sichern und zu einem späteren Zeitpunkt weiter bearbeiten.

Mit dem Speichern einer Sitzung als StatFolio werden die Definitionen der Analysen gesichert, nicht jedoch die Ergebnis-Ausgabe. Beim Öffnen eines StatFolio werden die Daten aus den eingebundenen Datenquellen neu gelesen und alle Analysen neu erstellt. Ein StatFolio stellt auf diese Weise ein einfaches Mittel zur Wiederholung von Analysen zu einem späteren Zeitpunkt und aus möglicherweise veränderten Daten dar.

Sie können zudem Scripte erzeugen, die bei jedem Öffnen eines StatFolio ausgeführt werden. Details zu den Scripten und weiteren Eigenschaften von StatFolios werden in diesem Kapitel beschrieben.

5.1 Speichern der Sitzung

Um den aktuellen Stand Ihrer STATGRAPHICS Centurion XVI-Sitzung zu sichern, wählen Sie im Hauptmenü *Datei – Speichern – StatFolio speichern*. In der aufgerufenen Dialogbox geben Sie dann einen Namen für das StatFolio ein:

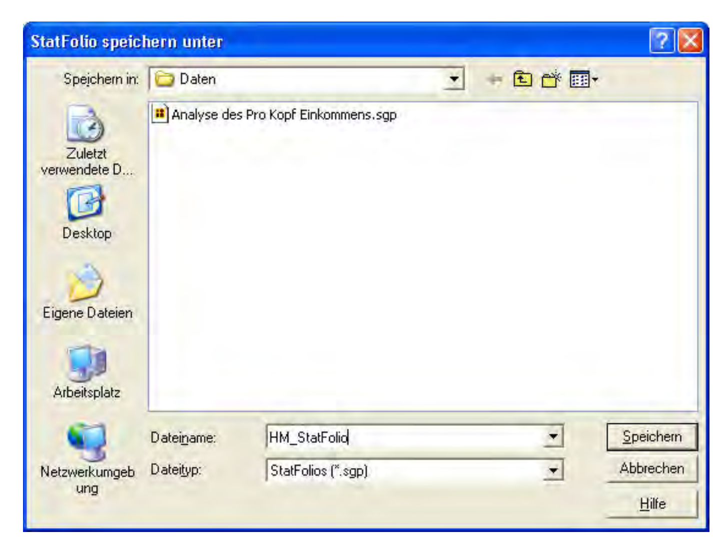


Abbildung 5.1: Dialogbox zum Speichern eines StatFolios

StatFolios werden als Dateien mit der Erweiterung *.sgp* gespeichert. Ein StatFolio enthält:

1. die Definition aller erstellten Analysen, einschließlich der Eingabe-Variablen, der Tabellen und Diagramme, der Einstellungen aller Optionen, der Veränderungen an den Diagrammen, etc.. Wenn ein StatFolio wieder geöffnet wird, werden die Analysen neu errechnet und alle Tabellen und Diagramme aktualisiert.

2. Verknüpfungen zu den im Datenbuch enthaltenen Datenquellen. Wenn die Daten in der Zeit zwischen dem Speichern und erneutem Öffnen eines StatFolio verändert werden, werden diese Veränderungen auch in der Analyse wiedergegeben.

3. Verknüpfungen zu einer StatGallery- und StatReporter-Datei, sofern dort Ergebnisse vor dem Speichern des StatFolio eingefügt wurden. Das Programm fragt beim Speichern des StatFolios auch die Namen für die StatGallery und den StatReporter ab.

5.2 Das StatFolio-Script

Bei dem erstmaligen Öffnen eines StatFolios wird der vorherige Stand aller Analyse-Fenster wiederhergestellt. Dann prüft STATGRAPHICS Centurion XVI, ob mit dem StatFolio ein Start-Script gespeichert wurde, das, wenn vorhanden, im Anschluss ausgeführt wird. Sie können ein Script erzeugen, indem Sie im *Bearbeiten*-Menü *StatFolio Start-Script* wählen. Es wird eine Dialogbox mit Eingabefeldern, über die eine Reihe von Aktionen definiert wird, aufgerufen:

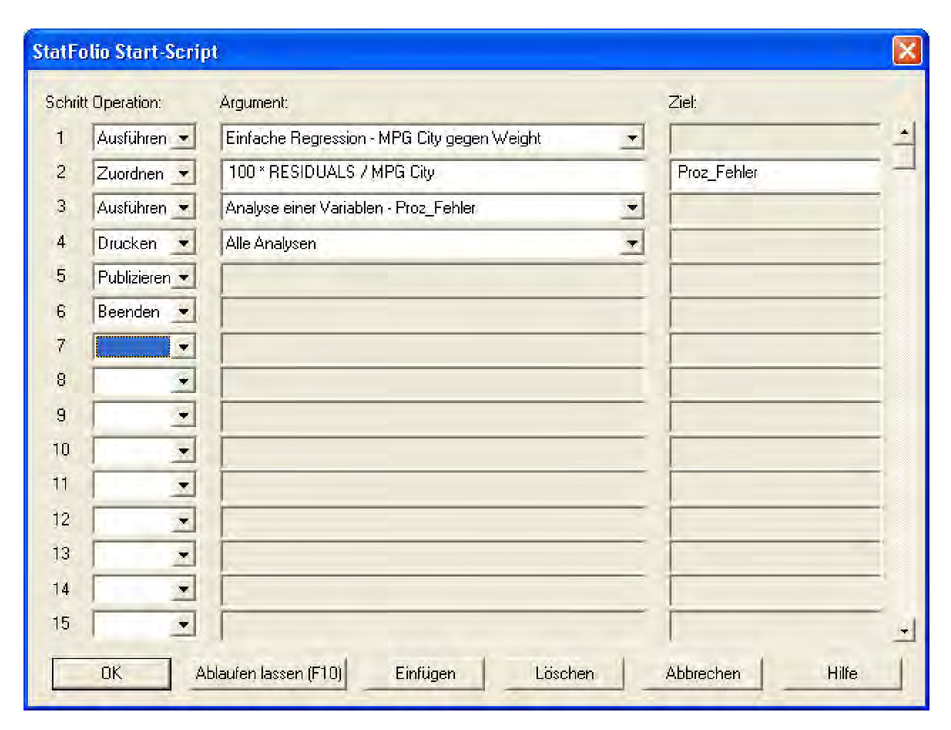


Abbildung 5.2: Start-Script-Dialogbox mit einem einfachen Start-Script für ein StatFolio

Die gewünschten Aktionen müssen hier in der Reihenfolge, in der sie ausgeführt werden sollen, eingegeben werden. Folgende Aktionen sind verfügbar:

Befehl	*Argument*	*Ziel*	*Beschreibung*
Execute (Ausführen)	Analyse-Titel		Aktualisiert die angegebene Analyse.
Assign (Zuweisen)	STATGRAPHICS Centurion-Ausdruck	Spaltenname	Berechnet den Ausdruck und weist ihn der angegebenen Spalte zu.
Print (Drucken)	Zu druckende(s) Fenster		Druckt die Inhalte des/der angegebenen Fenster(s).
Publish (Publizieren)			Ruft StatPublish auf und exportiert die Inhalte des StatFolios im HTML-Format.
Shell	Auszuführender Windows-Befehl	Befehl Argument	Bewirkt, dass Windows einen Befehl ausführt.
Delay (Verzögern)	Anzahl an Sekunden		Pausiert für die angegebene Zeit.
Load (Laden)	Name des StatFolio		Legt das nach dem Script-Lauf zu ladende StatFolio fest. Dies ermöglicht, dass mehrere StatFolios nacheinander ausgeführt werden können.
Exit (Beenden)			Beendet STATGRAPHICS Centurion XVI

Abbildung 5.3: Operationen für Start-Scripte

Im Beispiel in *Abbildung 5.2* wird eine *Einfache Regression* durchgeführt. Die Analyse geht davon aus, dass *Ergebnisse speichern* so eingestellt ist, dass die Residuen des angepassten Modells automatisch in der Spalte *RESIDUALS* gesichert werden. Die Residuen werden anschließend durch die ursprünglichen Datenwerte dividiert und mit 100 multipliziert, um die prozentualen Fehler zu berechnen, die der neuen Variable *Proz_Fehler* zugewiesen werden können. Die Werte in *Proz_Fehler* werden mit der Prozedur *Analyse einer Variablen* ausgewertet. Die Ergebnisse beider Analysen werden abschließend gedruckt und im Netz publiziert.

HINWEIS: StatFolios können über den Operator LADEN in einem Script verkettet werden, um eine weiteres StatFolio zu laden und dessen Script auszuführen. Zudem können Sie STATGRAPHICS Centurion XVI über den Operator BEENDEN automatisch schließen.

ANMERKUNG: Sie können die Ausführung eines Scripts unterdrücken, indem Sie im Register *Allgemein* der Dialogbox *Voreinstellungen*, die über das Menü *Bearbeiten* aufgerufen wird, das Kontrollkästchen *Start-Script deaktivieren* anklicken:

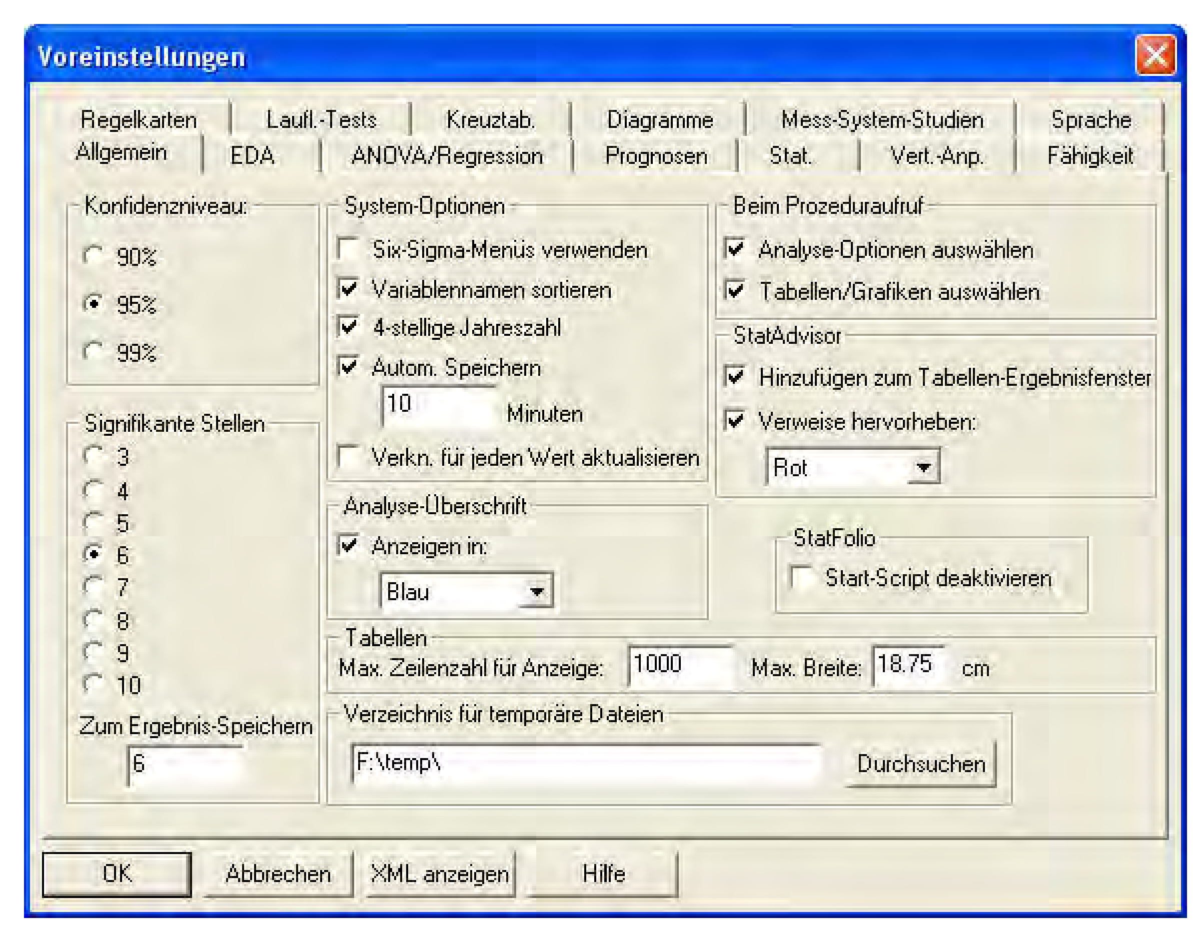


Abbildung 5.4: Start-Script deaktivieren

5.3 Abfragen von Datenquellen

Sobald ein StatFolio mit mehreren Analysen erstellt wurde, können die Daten aus den eingebundenen Datenquellen in festgelegten Zeitintervallen gelesen und die Analysen regelmäßig aktualisiert werden. Solch eine Funktion kann in der Dialogbox *Datenbuch-Eigenschaften* im *Bearbeiten*-Menü oder über *StatLink* im *Datei*-Menü eingestellt werden:

Abbildung 5.5: Dialogbox Datenbuch-Eigenschaften mit der Option, Datenquellen automatisch abzufragen

Um Datenquellen wiederholt abzufragen, führen Sie folgende Schritte durch:

1. Aktivieren Sie das Kontrollkästchen *Abfragen* für jede der automatisch abzufragenden Datenquellen.

2. Aktivieren Sie den Auswahlschalter *Ein* im Gruppenfeld *StatLink abfragen.*

3. Bestimmen Sie im Eingabefeld *Aktualisieren alle* ein Zeitintervall für die Abfrage der Datenquellen.

4. Aktivieren Sie das Kontrollkästchen *Script laufen lassen*, wenn das Start-Script nach jeder Abfrage automatisch ausgeführt werden soll.

Sofern das Start-Script einen *Publizieren*-Befehl enthält, lädt STATGRAPHICS Centurion XVI die Analyse-Ergebnisse automatisch auf einen Netzwerk-Server.

5.4 Publizieren der Daten im HTML-Format

Die Analyse-Ergebnisse in einem StatFolio können im HTML-Format publiziert werden, so dass man zum Anschauen nur den Standard-Webbrowser benötigt. Hierfür wählen Sie *StatPublish* im *Datei*-Menü. Es wird eine Dialogbox angezeigt, in der Sie angeben, welche Ergebnisse publiziert und wo sie abgelegt werden sollen:



Abbildung 5.6: StatPublish-Dialogbox zur Erstellung von HTML-Dateien

In den Feldern dieser Dialogbox legen Sie folgendes fest:

- **HTML-Datei in einem lokalen Verzeichnis:** Name der HTML-Datei , die das Inhaltsverzeichnis für das StatFolio enthalten soll. In der Datei werden die Inhalte des StatFolio aufgelistet und Verknüpfungen zu weiteren HTML-Dateien mit den einzelnen Fenstern des StatFolio angezeigt. Die Datei wird standardmäßig in das gleiche Verzeichnis, in dem sich auch das StatFolio befindet, und unter dem gleichen Namen abgelegt. Der Name besitzt jedoch die Erweiterung *.htm* anstelle von *.sgp*. Das publizierte StatFolio, d.h. die neue HTML-Datei, wird standardmäßig in einem Webbrowser geöffnet.

- **URL der FTP-Seite:** Die als HTML-Dateien publizierten Ergebnisse werden zunächst in das oben angegebene Heimverzeichnis abgelegt. Sie enthalten die HTML-Dateien, Bilddateien mit den Diagrammen und weitere Dateien. Wenn das Eingabefeld *URL der FTP-Seite* einen Eintrag enthält, werden alle diese Dateien zudem mit der angegebenen URL hochgeladen. Als URL wird in der Regel ein Verzeichnis auf einem Server eingetragen. Für die URL müssen Sie einen FTP-Schreibzugang haben – dies muss möglicherweise vorher von Ihrem Netzwerk-Administrator eingerichtet werden.

- **FTP-Benutzername:** Benutzername für den FTP-Zugang zu der angegebenen URL.

- **FTP-Passwort:** Passwort für den FTP-Zugang zu der angegebenen URL.

- **Inhalte:** Aktivieren Sie die Kontrollkästchen aller StatFolio-Fenster, die Sie publizieren möchten.

- **Grafikhöhe und –breite in Pixel:** Gibt die Größe der ins HTML-Format einzubettenden Diagramme an.

- **Bildformat:** Die Diagramme können in den folgenden drei Formaten in eine HTML-Datei eingebettet werden:

    1. *JPEG* – statische Bilder im JPEG-Format. Es werden Dateien mit Namen wie „pubexample_analysis1_graph1.jpg" erzeugt.

    2. *PNG* – statische Bilder im PNG-Format. Es werden Dateien mit Namen wie „pubexample_analysis1_graph1.png" erzeugt.

    3. *Java-Applets* – dynamische Ausgaben, die aktualisiert werden können, während sie im Browser angezeigt sind. Während der Anzeige im Browser wird das Diagramm über eine Hilfsdatei mit einem Namen wie „pubexample_analysis1_graph1.sgz" aktualisiert. Diese

Option wurde für eine gemeinsame Ausführung mit der Echtzeit-Abfrage von Datenquellen über StatLink entworfen. Die Details finden Sie im PDF-Dokument *Dynamic Data Processing and Analysis*. HINWEIS: Mit dieser Option können nicht alle Diagramme einwandfrei publiziert werden. Sollten eins oder mehrere Diagramme in der Publikation nicht richtig angezeigt werden, wählen Sie bitte eine andere Option.

- **Interaktive Applets:** Mit dieser Funktion können die Informationen zu den Datenwerten der Diagramme, die als Applets publiziert sind, angezeigt werden. Wird im Browser ein Punkt des Diagramms markiert, so werden die entsprechenden Datenwerte angezeigt.

Nach dem Ausfüllen der Eingabefelder wählen Sie *OK*, um das StatFolio zu publizieren.

Um ein publiziertes StatFolio anzusehen, starten Sie einen Webbrowser und öffnen über das Datei-Menü des Browsers die im ersten Eingabefeld von *Abbildung 5.6* angegebene Adresse. Sie können sich die Ergebnisse außerdem über *Publizierte Ergebnisse ansehen* im STATGRAPHICS Centurion XVI-*Datei*-Menü ansehen.

ANMERKUNG: Die Namen der Tabellen und Diagramme, die in HTML-Dateien eingebettet werden, werden von StatPublish automatisch generiert. Im Webbrowser können Sie den HTML-Quelltext anzeigen, um die Dateinamen herauszufinden. Wenn gewünscht, können Sie die Dateien in dieser Form auch in Ihre Website einbinden.

Kapitel

6

StatGallery

Neben- und Übereinanderlegen mehrerer Diagramme.

Die StatGallery ist ein besonderes Fenster in STATGRAPHICS Centurion XVI, in dem Diagramme, die in unterschiedlichen Prozeduren erzeugt wurden, einander gegenübergestellt oder übereinandergelegt werden können. Durch das Nebeneinanderlegen können zwei Gruppen von Daten, zwei statistische Modelle oder zwei Niveaus eines Konturen-Diagramms besonders effektiv miteinander verglichen werden. Das Übereinanderlegen der Diagramme erzeugt einzigartige Darstellungen, die an keiner anderen Stelle des Programms generiert werden können.

StatGallery-Ergebnisse werden in Dateien mit der Erweiterung *.sgg* gesichert. Wenn Sie Ergebnisse in der StatGallery ablegen, wird in das aktuelle StatFolio ein Verweis auf die StatGallery-Datei eingefügt. Wird das StatFolio dann zu einem späteren Zeitpunkt wieder geöffnet, so wird automatisch auch die zugehörige StatGallery geladen.

6.1 Konfigurieren eines StatGallery-Fensters

Die StatGallery wird mit dem Start von STATGRAPHICS Centurion XVI in einem separaten Fenster geöffnet. Das Fenster umfasst eine oder mehrere Seiten, von denen jede bis zu neun Diagramme enthalten kann. Standardmäßig ist jede Seite der Gallery so eingestellt, dass sie vier Diagramme anzeigen kann:

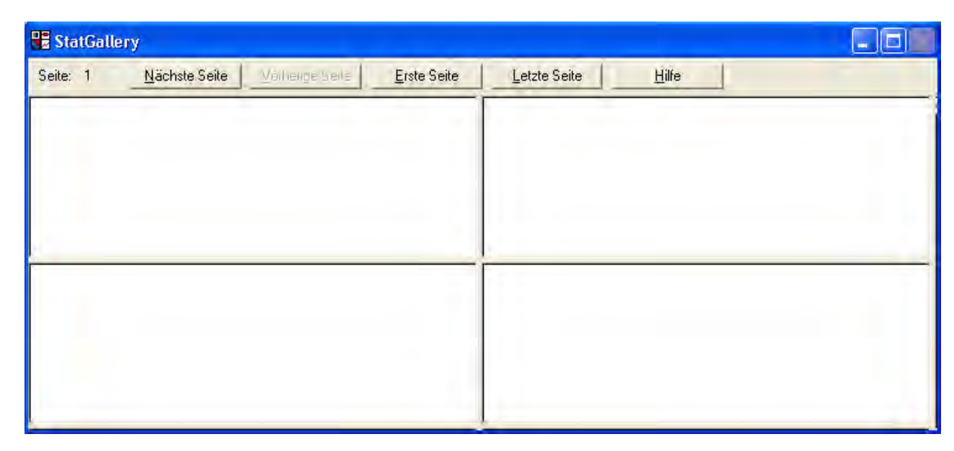


Abbildung 6.1: Das StatGallery-Fenster

Mit den Schaltflächen über den vier Ausschnitten (Ergebnisfenster) des Fensters kann zwischen den Seiten der Gallery navigiert werden. Wenn Sie die Anzahl der auf einer Seite enthaltenen Diagramme ändern möchten, öffnen Sie über die linke Maustaste das Kontextmenü der Gallery und wählen *Ergebnisfenster anordnen* aus. Sie können bis zu neun Diagramme pro Seite einstellen:



Abbildung 6.2: Alternative StatGallery-Seitenkonfigurationen

Die *Anordnung*en links im Gruppenfeld der Dialogbox entsprechen einer Matrix aus Zeilen und Spalten. Die Option *Spaltenweise* rechts ermöglicht, eine Anordnung mit unterschiedlichen Anzahlen an Zeilen in jeder der bis zu drei Spalten zu definieren.

Sie können auch direkt die beweglichen Trennbalken im StatGallery-Fenster verwenden, um die Ergebnisfenster in die gewünschte Anordnung zu bringen.

6.2 Einfügen der Diagramme in die StatGallery

Um ein Diagramm in die StatGallery einzufügen, kopieren Sie es zunächst aus dem Analyse-Fenster, in dem es erstellt wurde, in die Windows-Zwischenablage. Wenn Sie z.B. Konturen-Diagramme, die mit der DOE-*Auswertung des Versuchsplans* erzeugt wurden, auf zwei verschiedenen Stufen eines gewählten Versuchsfaktors darstellen wollen, nehmen Sie folgende Schritte vor:

1. Konfigurieren Sie eine Seite der StatGallery so, dass die Diagramme *Links und rechts* angezeigt werden.

2. Erzeugen Sie mit der *Auswertung des Versuchsplans* ein Konturen-Diagramm für eine Stufe des Versuchsfaktors und kopieren Sie es in die Windows-Zwischenablage.

3. Aktivieren Sie das StatGallery-Fenster. Rufen Sie das Kontextmenü des linken Ergebnisfensters über die rechte Maustaste auf und wählen Sie *Einfügen*, um das Konturen-Diagramm einzusetzen.

4. Zurück im Fenster der *Auswertung des Versuchsplans* erzeugen Sie ein zweites Konturen-Diagramm für eine andere Stufe des Versuchsfaktors. Kopieren Sie das Diagramm in die Windows-Zwischenablage.

5. Kehren Sie nun zurück in das StatGallery-Fenster. Rufen Sie das Kontextmenü des rechten Ergebnisfensters auf und wählen Sie *Einfügen*. Hiermit wird das zweite Konturen-Diagramm in der StatGallery neben dem ersten eingesetzt.

Das StatGallery-Fenster sollte nun folgendermaßen aussehen:

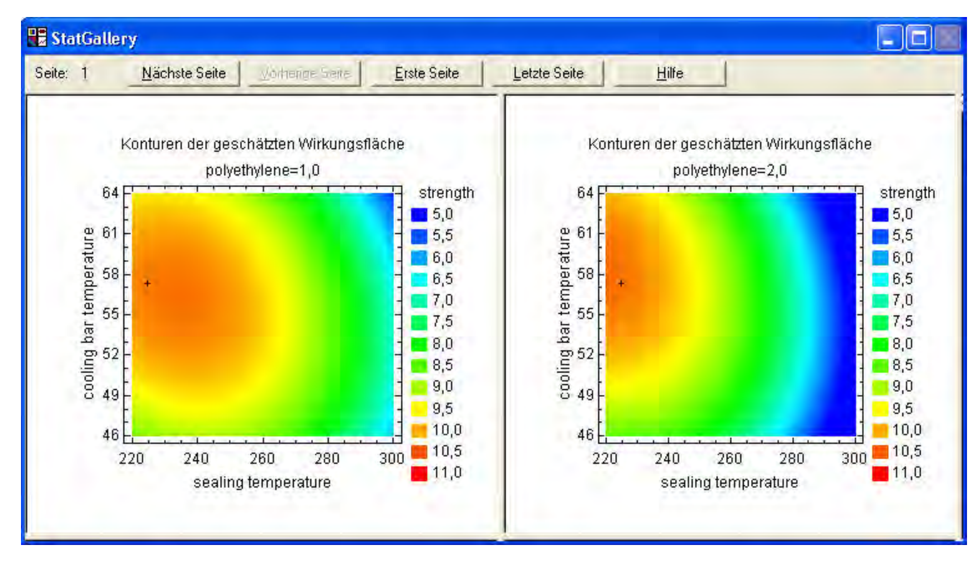


Abbildung 6.3: Gegenübergestellte Diagramme in der StatGallery

Durch den Vergleich der Diagramme wird deutlich, dass die Zielgröße *strength* mit zunehmendem Polyethylengehalt kleiner wird.

Um ein Diagramm in die StatGallery einfügen, können Sie im Kontextmenü des Gallery-Fensters auch *Einfügen und verknüpfen* anstelle von *Einfügen* auswählen. Mit *Einfügen und verknüpfen* wird das Diagramm in der Gallery mit dem Analyse-Fenster, in dem es erzeugt wurde, verknüpft, so dass nach einer Veränderung des Analyse-Fensters auch die Anzeige in der StatGallery aktualisiert wird.

6.3 Übereinanderlegen von Diagrammen

Wird ein Diagramm in ein Ergebnisfenster der StatGallery eingesetzt, das bereits ein Diagramm enthält, können Sie entscheiden, ob das bereits vorhandene Diagramm durch das einzufügende ersetzt oder ob es über dem vorhandenen platziert werden soll. Das Übereinanderlegen von zwei Diagrammen bietet sich an, wenn z. B. zwei statistische Modelle angepasst werden sollen:

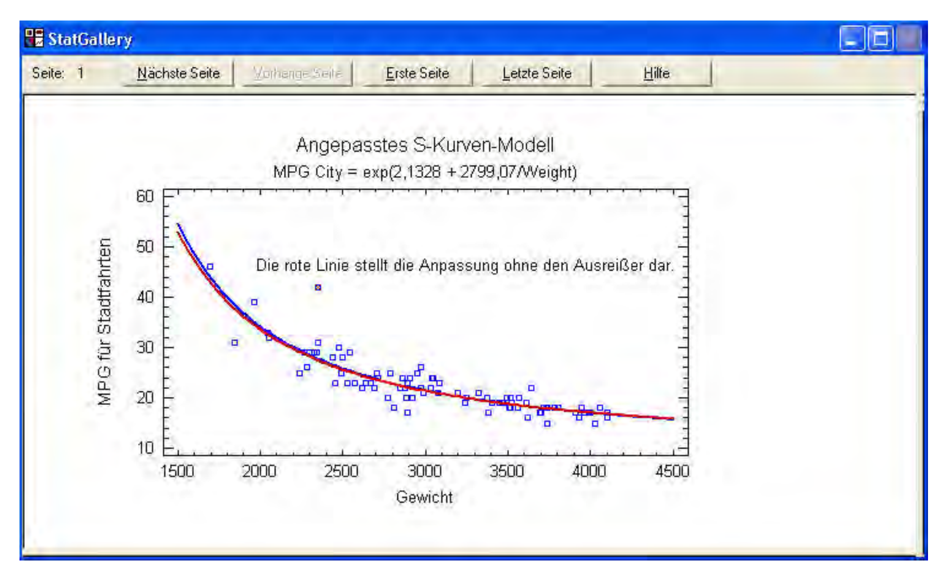


Abbildung 6.4: Übereinandergelegte Diagramme in der StatGallery

Wenn ein Diagramm über ein weiteres, das sich bereits in der StatGallery befindet, gelegt wird, werden vom zweiten Diagramm ausschließlich die Inhalte *innerhalb* der Achsen der Anzeige eingesetzt. Die Texte und Beschriftungen des zweiten Diagramms werden **nicht** angezeigt.

> ANMERKUNG: Wenn die Skalierung des zweiten Diagramms sich von dem ersten unterscheidet, so wird das zweite Diagramm an die Skalierung des ersten angepasst.

6.4 Bearbeiten eines Diagramms in der StatGallery

Nachdem ein Diagramm in die StatGallery eingefügt wurde, können noch einige Aspekte geändert werden.

6.4.1 Hinzufügen von Elementen

Um ein Element in ein Diagramm einzufügen:

1. Maximieren Sie das Diagramm über einen Doppelklick in das Ergebnisfenster.

2. Rufen Sie das Kontextmenü des Diagramms über die rechte Maustaste auf und wählen Sie *Element hinzufügen*. Der folgende Dialog wird geöffnet:



Abbildung 6.5: Dialogbox zum Hinzufügen eines Elements

3. Wählen Sie hier das Element, das Sie dem Diagramm hinzufügen möchten, aus.

Die ersten fünf Elemente in der Dialogbox in *Abbildung 6.5* können direkt eingefügt werden, indem Sie die entsprechende Schaltfläche auswählen, im Diagramm die Maus an die gewünschte Position setzen und das Element mit gedrückter Maustaste auf die gewünschte Größe ziehen. Die letzte Schaltfläche aktiviert den Textmodus und ruft eine Texteingabe-Dialogbox auf, sobald Sie das Diagramm anklicken. Der hier eingegebene Text kann mit der Maus an jede beliebige Position bewegt werden.

6.4.2 Bearbeiten der Elemente

Um ein Element in der StatGallery zu bearbeiten:

1. Maximieren Sie das Diagramm über einen Doppelklick in das Ergebnisfenster.

2. Markieren Sie das zu bearbeitende Element. Ein markiertes Element erkennen Sie an den kleinen rechteckigen Kästchen, welche die äußere Form des Elements umgeben.

3. Rufen Sie das Kontextmenü des Diagramms über die rechte Maustaste auf und wählen Sie *Element bearbeiten*.

Es wird eine dem Elementtyp angepasste Dialogbox geöffnet, über die Sie die Änderungen vornehmen können.

6.4.3 Löschen der Elemente

Um ein Element in der StatGallery zu löschen:

1. Maximieren Sie das Diagramm über einen Doppelklick in das Ergebnisfenster.

2. Markieren Sie das zu löschende Element mit der Maustaste.

3. Rufen Sie das Kontextmenü des Diagramms über die rechte Maustaste auf und wählen Sie *Element löschen*.

6.5 Drucken der StatGallery

Um die Elemente in der StatGallery zu drucken:

1. Aktivieren Sie das StatGallery-Fenster, indem Sie es mit der Maus anklicken.

2. Wählen Sie das *Drucken*-Symbol auf der Symbolleiste aus oder, alternativ, rufen Sie das Kontextmenü des Fensters auf und wählen hier *Drucken*.

Sie können eine einzige, mehrere oder alle Seiten der StatGallery drucken.

Kapitel 7

StatReporter

Einfügen der Analysen in den StatReporter, Kommentieren der Ergebnisse, Sichern der Ergebnisse im RTF-Format für den Import in Microsoft Word.

Der StatReporter ist ein Fenster, über das die Ergebnisse der verschiedenen statistischen Prozeduren in einen formellen Bericht integriert werden können. Es ist eine autonome Version von WordPad, die in STATGRAPHICS Centurion XVI ausgeführt werden kann. Der StatReporter ermöglicht die folgenden Funktionen:

1. Erstellen eines vollständigen Berichts in STATGRAPHICS Centurion XVI, ohne dass eine weitere Anwendung benötigt wird. Dies ist vor allem in Unternehmensbereichen mit eingeschränkten Ressourcen, wie z. B. Produktionsstätten, hilfreich.

2. Sichern der Inhalte des StatReporter in einer RTF-Datei (Rich Text Format), die von anderen Programmen wie Microsoft Word direkt gelesen werden kann.

7.1 Das StatReporter-Fenster

Der StatReporter ist ein separates Fenster in STATGRAPHICS Centurion XVI, das automatisch beim Starten des Programms erstellt wird. Das Fenster besitzt eine eigene Symbolleiste mit Formatierungswerkzeugen für die eingegebenen Texte:

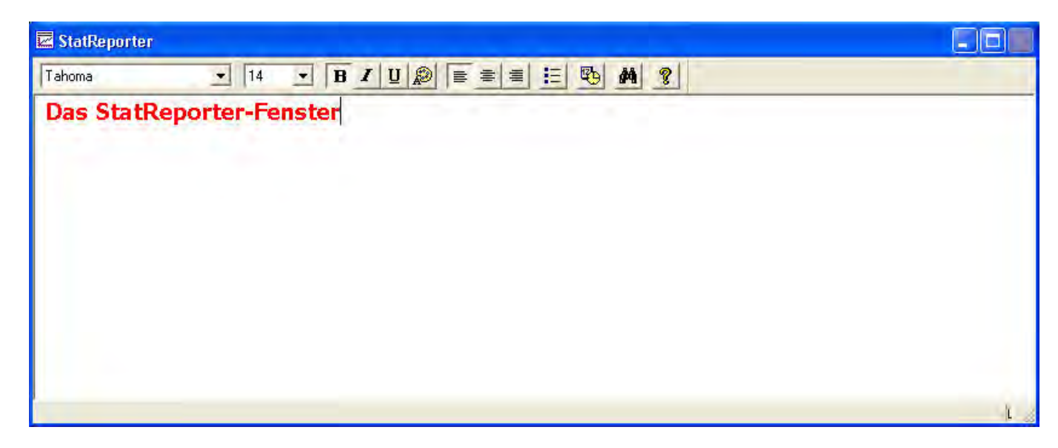


Abbildung 7.1: Das StatReporter-Fenster

In das Fenster können Sie Text eingeben oder Ergebnisse einfügen, die in STATGRAPHICS Centurion XVI erzeugt wurden.

7.2 Einfügen der Ergebnisse in den StatReporter

Mit STATGRAPHICS Centurion XVI können die Ergebnisse in drei unterschiedlichen Methoden in den StatReporter eingefügt werden:

1. Um eine einzige Tabelle oder ein Diagramm in den StatReporter einzufügen, maximieren Sie zunächst das gewünschte Ergebnisfenster und wählen *Kopieren* im *Datei*-Menü, um die Inhalte in die Windows-Zwischenablage zu kopieren. Begeben Sie sich dann in das StatReporter-Fenster, markieren Sie mit der Maus die gewünschte Einfüge-Position und wählen Sie *Datei – Einfügen*.

2. Alternativ maximieren Sie das Ergebnisfenster mit der gewünschten Tabelle oder dem Diagramm und rufen über die rechte Maustaste das Kontextmenü auf. Hier wählen Sie *Ergebnisfenster in StatReporter kopieren*. Auf diesem Weg wird die Tabelle oder das Diagramm automatisch an der Stelle in den StatReporter eingefügt, an der sich der Mauszeiger gerade befindet.

3. Um alle Ergebnisse eines Analyse-Fensters einzufügen, rufen Sie das Kontextmenü des Fensters auf und wählen *Analyse in StatReporter kopieren*. So werden alle Tabellen und Diagramme des Analyse-Fensters in den StatReporter eingesetzt.

Die eingefügten Inhalte sind statisch, d. h. im StatReporter verändern (z.B. aktualisieren) sie sich nicht. Sie können jedoch eine Tabelle oder ein Diagramm im StatReporter mit der Quelle

verknüpfen, indem Sie sie/es auf die unter 1. beschriebene Weise einfügen, jedoch anstelle von *Einfügen* die Option *Verknüpfung einfügen* wählen. Eine so in den StatReporter eingesetzte Tabelle oder Diagramm wird automatisch aktualisiert, wenn sie in dem Fenster, aus dem sie eingefügt wurde, verändert wird.

7.3 Bearbeiten der Inhalte des StatReporters

Über die Symbolleiste des StatReporters können die in das Fenster eingefügten Inhalte bearbeitet werden. Um Text zu ändern, markieren Sie den entsprechenden Text und wählen die gewünschte Schaltfläche auf der StatReporter-Symbolleiste aus. Zudem können Sie über die Schaltfläche *Datum/Zeit* das aktuelle Datum und die Zeit in das Fensters einfügen.

7.4 Speichern des StatReporters

Die Inhalte des StatReporters werden über *Datei – Speichern – StatReporter speichern* im Hauptmenü gesichert. Hier geben Sie auch einen Dateinamen an. Der StatReporter wird als RTF-Datei gespeichert, die direkt von Programmen wie Microsoft Word gelesen werden kann.

Wann immer Sie ein StatFolio öffnen, wird nun auch automatisch der StatReporter geladen, der beim Speichern des StatFolio vorhanden war. Über *Datei – Öffnen* kann ein StatReporter aber auch separat geöffnet werden.

Kapitel

8

StatWizard

Auswählen passender statistischer Analysen, Suchen nach den benötigten Statistiken und Tests und Erzeugen von Mehrfach-Fenstern entsprechend den Faktorstufen.

Der StatWizard ist ein besonderes Feature von STATGRAPHICS Centurion XVI, das Ihre Arbeiten auf vielfältige Weise unterstützt:

1. Er hilft beim Erzeugen eines neuen Datenblatts oder beim Einlesen bestehender Datenquellen in ein Datenblatt.

2. Er schlägt Analysen für die Daten vor auf Basis des Datentyps.

3. Er sucht nach benötigten Statistiken oder Tests und führt Sie zu den Prozeduren, die diese berechnen.

4. Er hilft bei der Definition von Daten-Transformationen oder bei der Auswahl von Untergruppen in den Daten.

5. Er kann gewünschte Analysen für jede Untergruppe in einer Datenspalte wiederholen.

Über die StatWizard-Schaltfläche auf der Symbolleiste können Sie jederzeit auf den StatWizard zugreifen.

8.1 Zugreifen auf Daten oder Erstellen einer neuen Studie

Falls der StatWizard aktiviert wird, wenn das Datenbuch leer ist , öffnet er eine Dialogbox, die Sie befragt, wie Sie mit Ihren Daten vorgehen möchten:

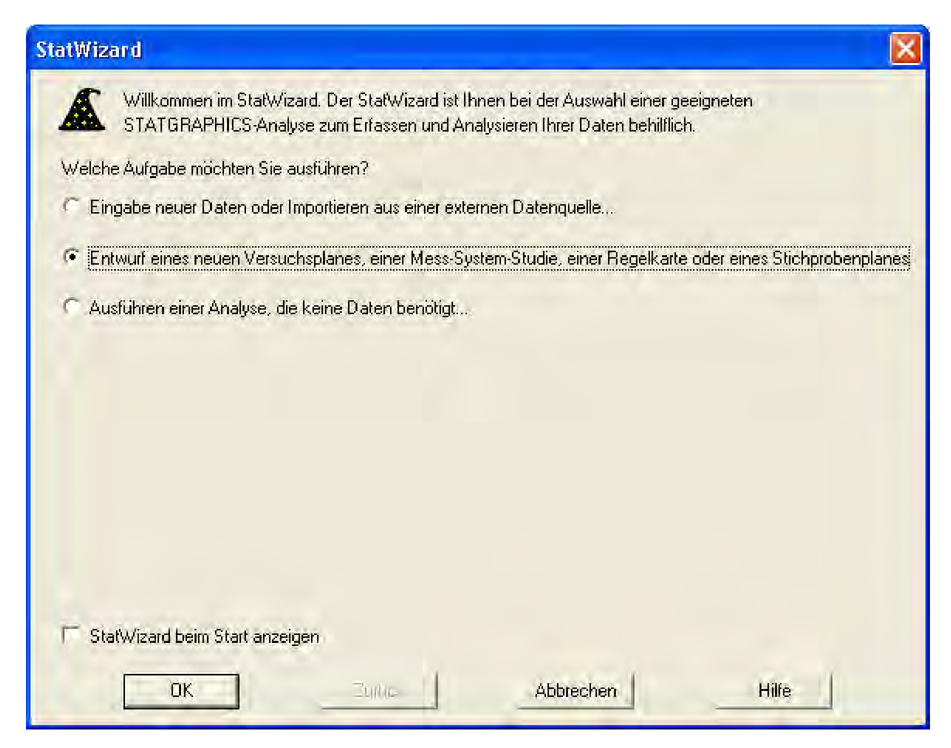


Abbildung 8.1: Dateneingabe-Dialogbox des StatWizard

Es werden drei Auswahlmöglichkeiten angeboten:

1. Sie möchten neue Daten in das STATGRAPHICS Centurion XVI-Datenbuch laden. Der Wizard wird Sie in diesem Fall – wie in den ersten Kapiteln beschrieben – durch eine Reihe weiterer Dialogboxen führen, um die Spalten des Datenblatts zu definieren oder eine Datenquelle einzubinden.

2. Sie möchten eine neue Studie entwerfen, bevor Sie Daten zusammenstellen. In diesem Fall wird der StatWizard Sie bitten, die Art der zu erstellenden Studie zu bestimmen, und Sie durch eine Reihe von Dialogboxen führen, über die Sie diese Studie definieren.

3. Sie möchten eine Analyse durchführen, für die keine Daten benötigt werden. In diesem Fall listet der StatWizard die möglichen Analysen auf, bittet Sie dann, eine auszuwählen, und führt Sie direkt zu dieser Analyse.

Angenommen, Sie möchten eine neue Messmittelfähigkeitsstudie einrichten, um die Wiederhol- und Vergleichspräzision (R&R) eines Messprozesses auswerten zu können. Wenn Sie das zweite Optionsfeld in *Abbildung 8.1* auswählen und die Dialogbox über *OK* verlassen, werden Ihnen anschließend die folgenden Optionen angezeigt:

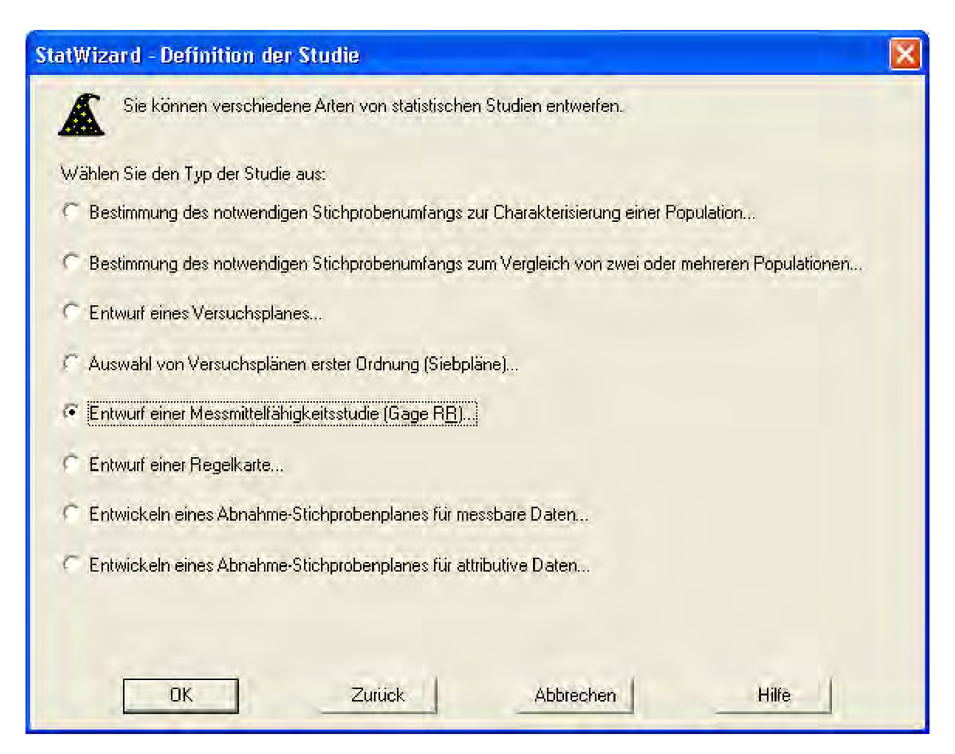


Abbildung 8.2: Dialogbox des StatWizard zur Definition einer Studie

Markieren Sie hier das Optionsfeld *Entwurf einer Messmittelfähigkeitsstudie (Gage R&R)* und wählen Sie *OK*, um die nächste Dialogbox aufzurufen, in der Sie Informationen zu der Studie eingeben:

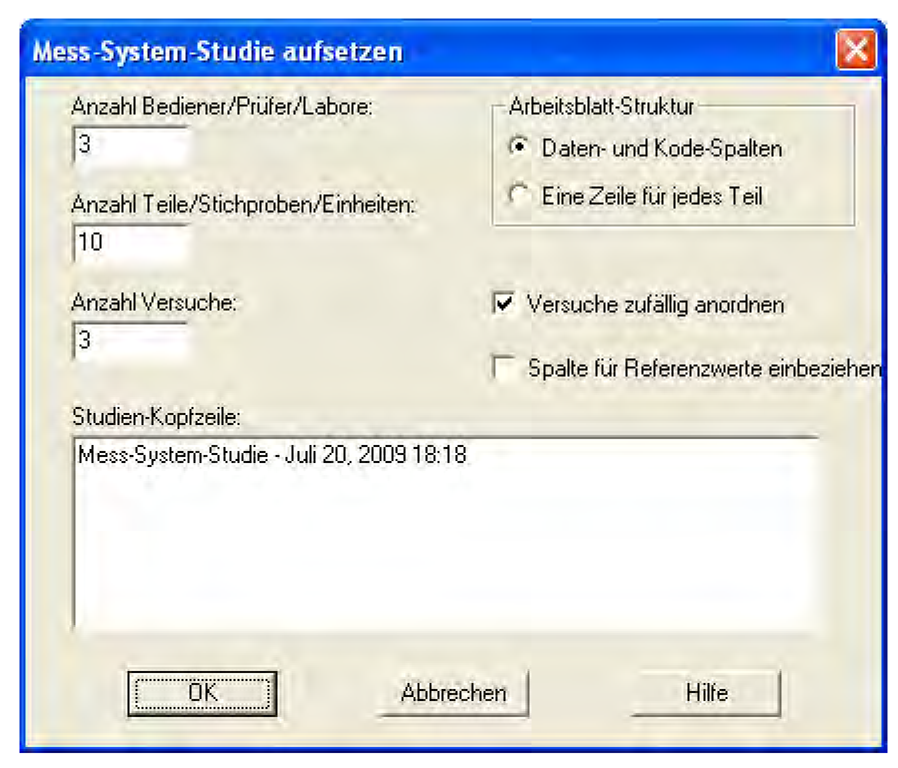


Abbildung 8.3: Dialogbox des StatWizard zum Einrichten einer Messmittelfähigkeitsstudie

Geben Sie in der Dialogbox die Anzahl der in der Studie involvierten Bediener, die Anzahl der zu messenden Teile und die Anzahl der Messungen, die jeder Bediener an jedem Teil durchführen wird, ein. Zudem können Sie in dem Eingabefeld *Studien-Kopfzeile* eine Überschrift für die Studie eingeben.

Die letzte Dialogbox fragt nach den Namen der Bediener, Prüfer oder Labore, welche die Messungen vornehmen:

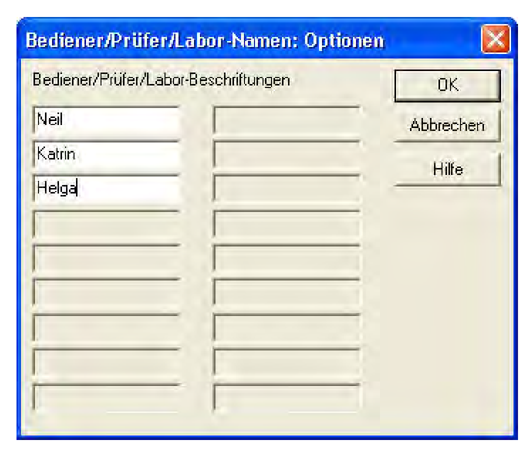


Abbildung 8.4: Dialogbox zum Eingeben der Namen der Ausführenden

Der StatWizard erstellt nun die gewünschte Studie und setzt sie in ein Datenblatt des Datenbuchs ein:

<unbezeichnet>

	Bediener	Teile	Versuche	Messungen	Kopfzeile
1	Neil	3	1		Mess-System-Studie - 31. Mai 2007
2	Neil	10	1		
3	Neil	4	1		
4	Neil	2	1		
5	Neil	9	1		
6	Neil	1	1		
7	Neil	5	1		
8	Neil	8	1		
9	Neil	6	1		
10	Neil	7	1		
11	Neil	9	2		
12	Neil	7	2		

A B C D E F G H I J

Abbildung 8.5: Mit dem StatWizard erstellte Messmittelstudie

Die Studie kann nun durchgeführt und die Messungen in das Datenblatt eingegeben werden. Anschließend kann der StatWizard erneut aufgerufen werden, um die passenden Analyse-Prozeduren auszuwählen (oder Sie wählen die Analysen direkt im Hauptmenü aus).

8.2 Auswählen der Analysen für Ihre Daten

Wenn sich schon Daten im Datenbuch befinden, wird über die Schaltfläche *StatWizard* die folgende Dialogbox geöffnet, in der Sie eine oder mehrere durchzuführende Analysen wählen können:

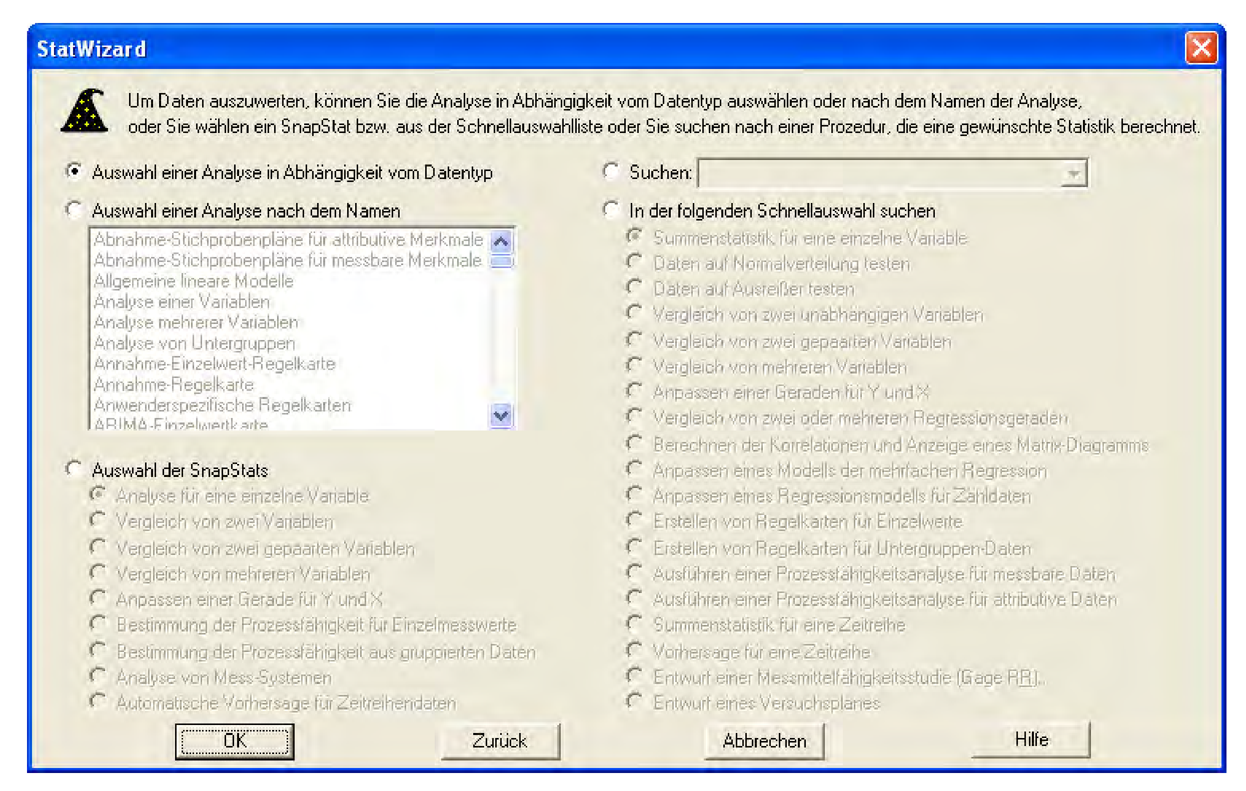


Abbildung 8.6: Dialogbox des StatWizard zur Auswahl der Analysen

Sie können nun zwischen den folgenden fünf Optionen wählen:

1. **Auswahl einer Analyse in Abhängigkeit vom Datentyp**: Zeigt eine weitere Dialogbox an, in der Sie Informationen zu den auszuwertenden Daten eingeben; anschließend werden alle relevanten Prozeduren aufgelistet.

2. **Auswahl einer Analyse nach dem Namen**: Zeigt alle verfügbaren Analysen in alphabetischer Reihenfolge an. Wählen Sie eine Analyse nach ihrem Namen aus und wählen Sie *OK*, um direkt zu der Dateneingabe-Dialogbox für diese Analyse zu gelangen. Dabei werden die üblichen Menüs übersprungen.

3. **Auswahl der SnapStats**: Ermöglicht die Auswahl eines SnapStat. SnapStats sind Überblicks-Analysen, die bereits vorformatierte Ergebnisse auf einer einzigen Seite erzeugen. Sie bieten zwar weniger Optionen als andere Analysen, sind jedoch sehr einfach zu erstellen.

4. **Suchen:** Öffnet eine Pulldown-Liste aller Statistiken, Tests, Grafiken und anderer Ergebnisse, die in STATGRAPHICS Centurion XVI erstellt werden können. Wenn Sie ein Element aus dieser Liste markieren, wird die Auswahl in dem Feld *Auswahl einer Analyse nach dem Namen* so geändert, dass nur noch die Analysen erscheinen, die das markierte Element errechnen können.

5. **In der folgenden Schnellauswahl suchen**: Listet einige der gebräuchlichsten Analysen auf. Wenn Sie hier eine Analyse und *OK* wählen, gelangen Sie direkt in die Dateneingabe-Dialogbox dieser Analyse.

Wenn Sie die Option 1. wählen, zeigt der StatWizard eine Dialogbox an, in der Sie die zu analysierenden Daten eingeben. Ist z. B. die *93cars.sgd*-Datei im Datenbuch geöffnet, sieht die Dialogbox wie folgt aus:

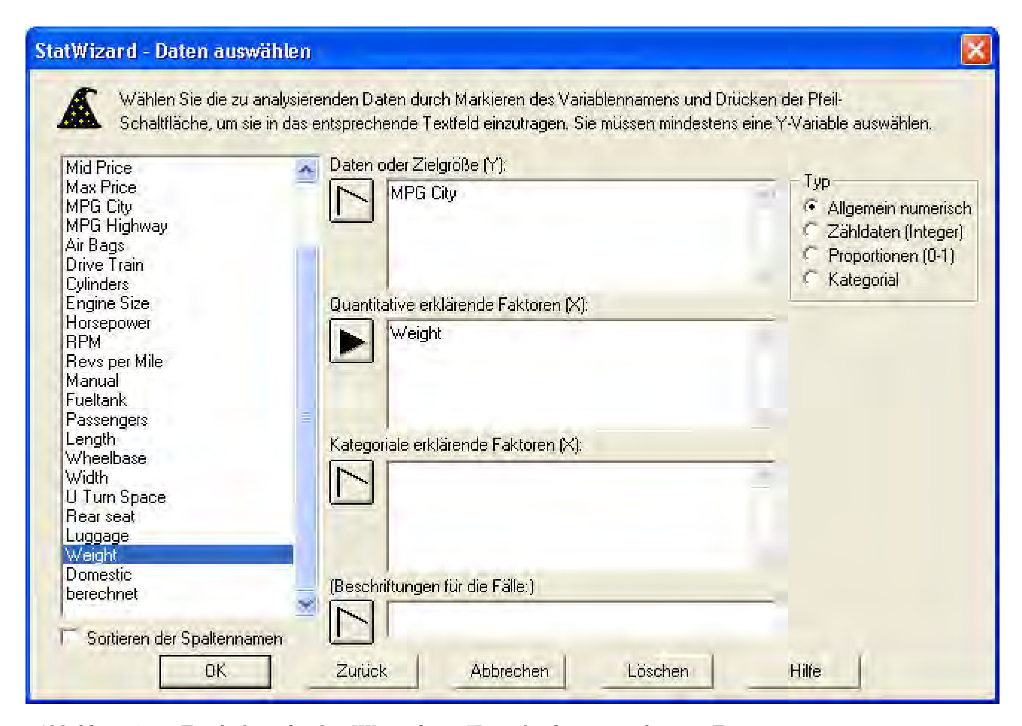


Abbildung 8.7: Dialogbox des StatWizard zur Eingabe der zu analysieren Daten

Die Dialogbox besitzt folgende Felder:

- *Daten oder Zielgröße (Y):* eine oder mehrere abhängige Variable, welche die zu analysierenden Werte enthalten. Wenn nur eine Spalte zu analysierende Daten enthält, dann muss sie hier eingegeben werden.

- *Typ:* der Typ der Daten, die in der/den abhängigen Variable(n) enthalten sind. Von dieser Eingabe hängt ab, welche Analysen die folgenden Dialogboxen anbieten.

- *Quantitative erklärende Faktoren (X)*: alle quantitativen Faktoren, die zur Vorhersage der abhängigen Variablen verwendet werden sollen. In einer Regression werden die unabhängigen Variablen hier eingegeben.

- *Kategoriale erklärende Faktoren (X)*: alle nicht-quantitativen Faktoren, die zur Vorhersage der abhängigen Variablen verwendet werden sollen. In einer ANOVA werden die unabhängigen Faktoren hier eingegeben.

- *Beschriftungen für die Fälle:* eine Spalte, welche die Beschriftungen für jede der Beobachtungen (Zeilen, Fälle) enthält.

Welche Prozeduren in den folgenden Dialogboxen angeboten werden, ist abhängig von den Einträgen, die in dieser Dialogbox gemacht werden.

In der nächsten Dialogbox geben Sie an, welche der Zeilen der Datei analysiert werden sollen:



Abbildung 8.8: Dialogbox des StatWizard zur Auswahl der zu analysieren Zeilen

Mit den ersten sechs Optionen wird vorausgesetzt, dass Sie nur eine einzige Analyse durchführen möchten. Mit der letzten Option werden mehrere Analyse-Fenster erstellt, d.h. eins für jeden Einzelwert in der angegebenen Spalte. So können Sie auf einfachem Weg eine BY-Variable für eine Reihe von Analysen festlegen.

Anschließend werden Sie gefragt, ob Sie eine der angegebenen Variablen transformieren möchten: Wenn Sie dies bejahen, wird die folgende Dialogbox angezeigt:

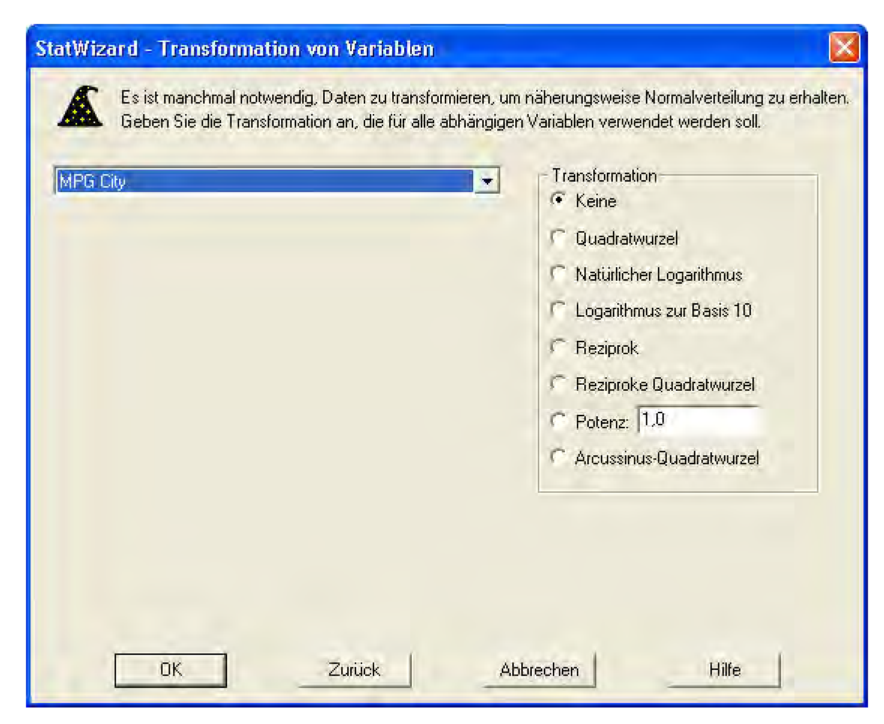


Abbildung 8.9: Dialogbox des StatWizard zum Transformieren von Variablen

Hier können Sie eine Transformation für eine oder mehrere Variablen auswählen. Für die gewählte Transformation wird dann der entsprechende Ausdruck generiert. Für die Quadratwurzel von *MPG City* würde z. B. der Ausdruck *SQRT(MPG City)* erstellt und in den Analyse-Prozeduren verwendet werden.

Abschließend wird eine Dialogbox geöffnet, in der alle geeigneten Analysen für den angegebenen Datentyp aufgelistet sind:

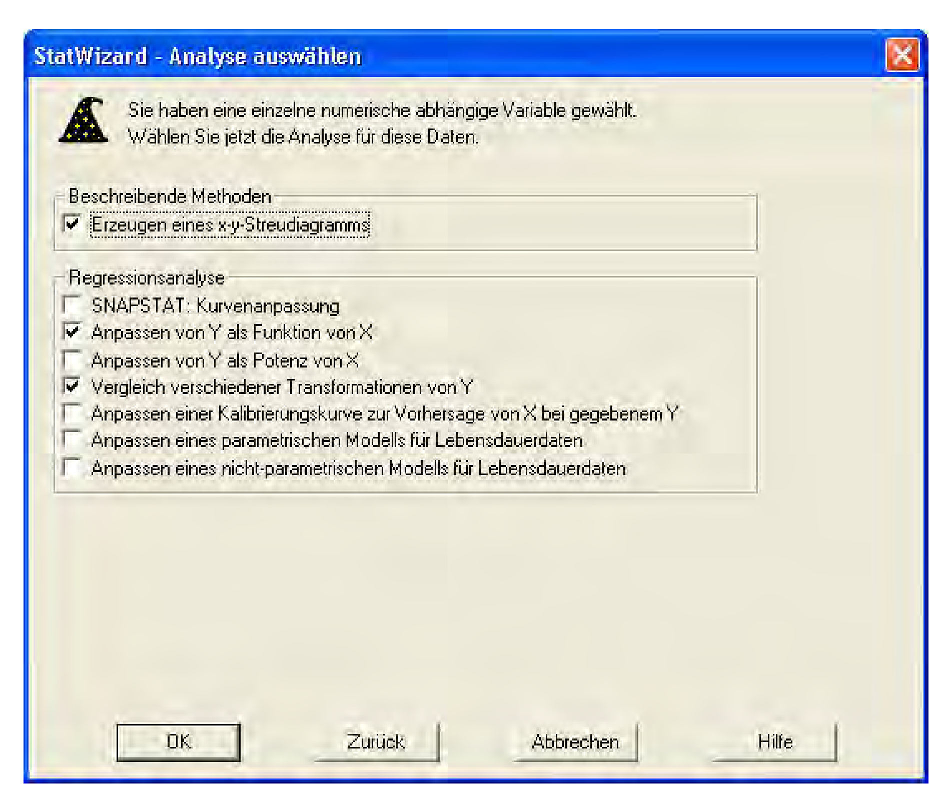


Abbildung 8.10: Dialogbox des StatWizard zur Auswahl einer Analyse

Hier wählen Sie eine oder mehrere Analysen aus. Nachdem Sie Ihre Angaben mit *OK* bestätigt haben, wird ein Analyse-Fenster für jede gewählte Analyse erstellt.

8.3 Suchen nach Statistiken oder Tests

Wenn Sie eine bestimmte Statistik oder einen Test berechnen möchten, aber unsicher sind, welche der Analysen diese Aufgabe erfüllt, dann können Sie Ihre Daten in ein Datenblatt eingeben und den *StatWizard* über die entsprechende Schaltfläche auf der Symbolleiste aufrufen. Wählen Sie in der Dialogbox des StatWizard *Suchen* und klicken Sie auf den Pfeil neben dem Eingabefeld, um die Pulldown-Liste mit den Suchbegriffen zu öffnen. Die Liste zeigt alle Statistiken, Tests und weitere Größen an, die STATGRAPHICS Centurion XVI berechnen kann:

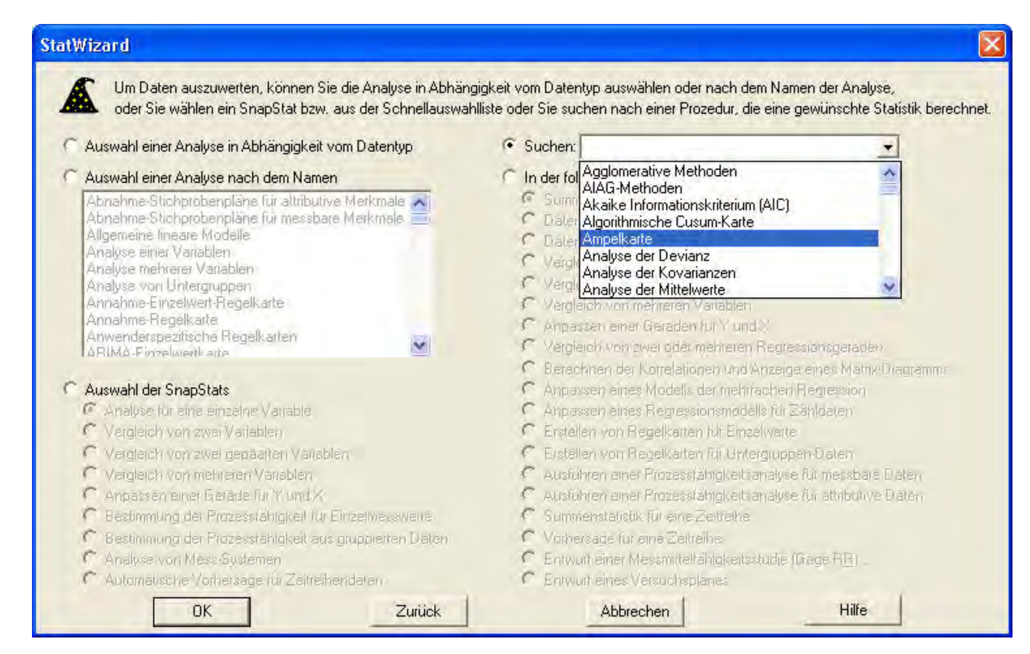


Abbildung 8.11: Verwenden der Suchen-Option des StatWizard

Wenn Sie ein Element in der Liste markieren, werden alle Analysen, mit denen es berechnet werden kann, in dem Feld *Auswahl einer Analyse nach dem Namen* angezeigt.

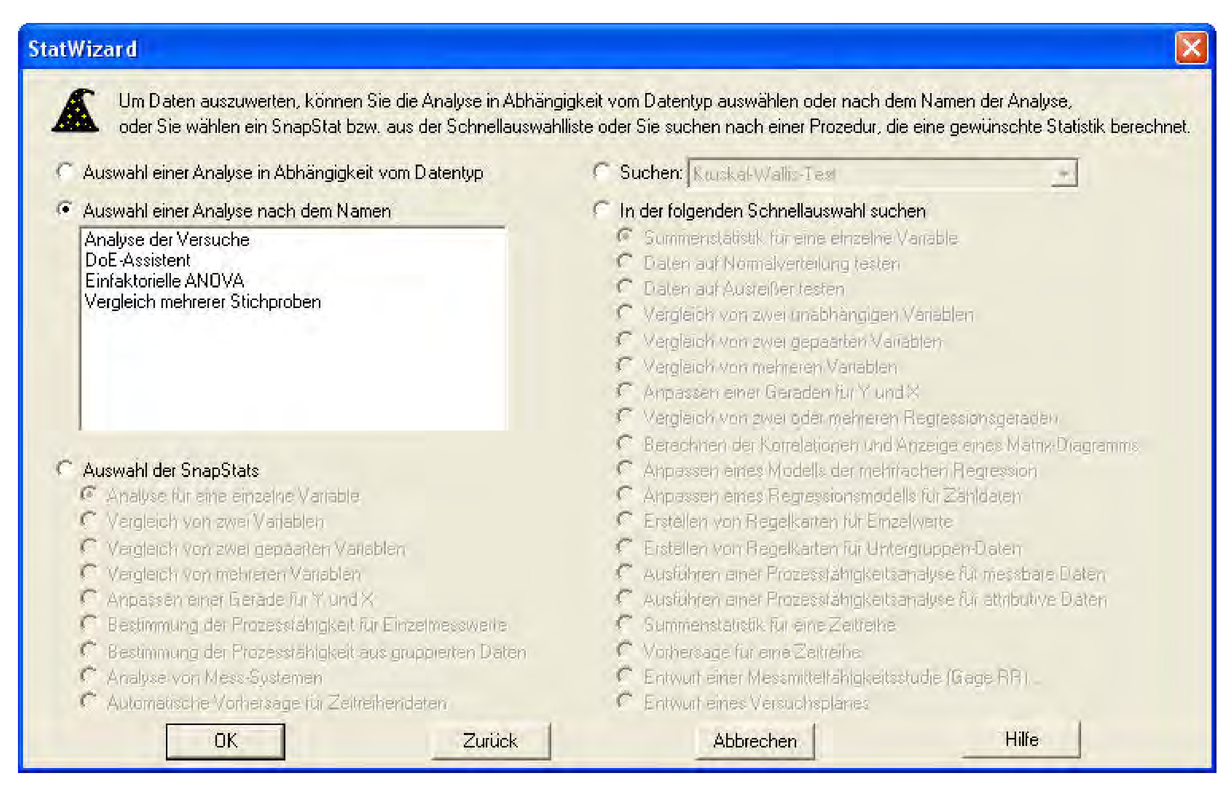


Abbildung 8.12: Liste aller Analysen, die dem Suchbegriff zugeordnet sind

Um hier eine Analyse auszuwählen und auszuführen, nehmen Sie folgende Schritte vor:

1. Markieren Sie das Optionsfeld *Auswahl einer Analyse nach dem Namen.*

2. Markieren Sie eine Analyse.

3. Wählen Sie *OK.*

Sie werden nun direkt zu der Dateneingabe-Dialogbox der gewählten Analyse geführt. Dabei werden die üblichen Menüs übersprungen.



Systemeinstellungen

Festlegen der Voreinstellungen des Systems.

Für jede der Optionen in STATGRAPHICS Centurion XVI ist bei ihrem erstmaligen Aufrufen ein Standard definiert, der an die Anforderungen der meisten Anwender angepasst ist. Wann immer nötig, können Sie neue Standardwerte für die Optionen einstellen. Hierfür gibt es im Programm drei Stellen:

1. **Allgemeine Systemführung:** wird in der Dialogbox *Voreinstellungen* festgelegt, die über das Menü *Bearbeiten* aufgerufen wird.

2. **Druckoptionen:** wird in der Dialogbox *Seite einrichten* festgelegt, die über das Menü *Datei* aufgerufen wird.

3. **Diagramme:** wird über *Diagramm-Optionen* eingestellt, während Sie ein Diagramm betrachten. Über das Register *Profile* in der *Diagramm-Optionen*-Dialogbox können Sie mehrere Zusammenstellungen von Diagramm-Eigenschaften (Profile) speichern.

9.1 Allgemeine Systemführung

Die Standardwerte für die allgemeine Systemführungen und bestimmte statistische Prozeduren können über *Voreinstellungen* im *Bearbeiten*-Menü geändert werden. Hierfür wird eine Dialogbox mit verschiedenen Registern – dem Register *Allgemein* für das allgemeine Systemverhalten und weiteren Registern für die statistischen Analysen – geöffnet:

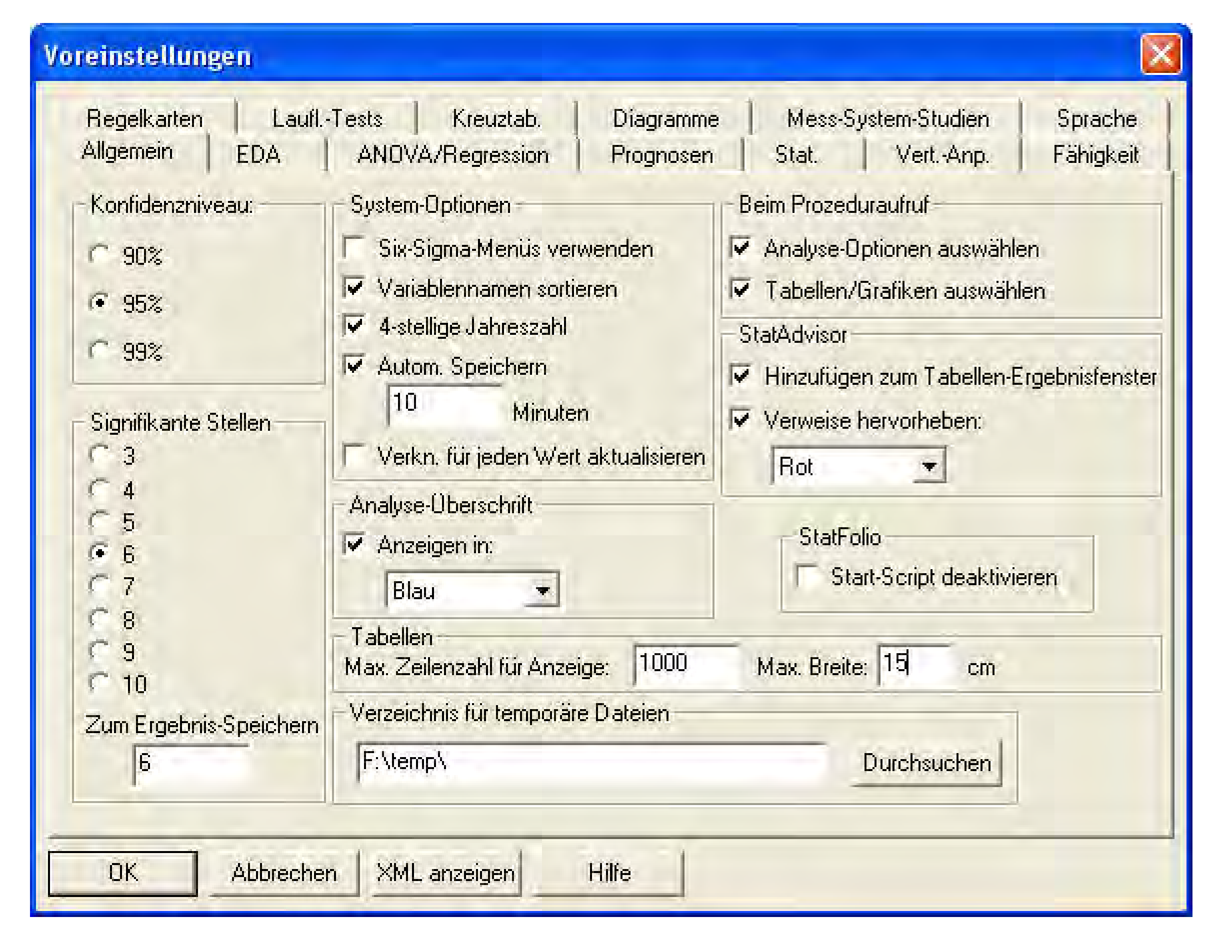


Abbildung 9.1: Voreinstellungen-Dialogbox

Zu den wichtigsten einstellbaren Optionen gehören:

- **Konfidenzniveau:** Standard-Prozentwert für Konfidenzgrenzen, Vorhersagegrenzen, Hypothesen-Tests und Interpretation der *p*-Werte durch den StatAdvisor.

- **Signifikante Stellen:** Anzahl der signifikanten Stellen bei der Ausgabe numerischer Ergebnisse. Es wird die angegebene Stellenzahl angezeigt; führende/endende Nullen werden hierbei unterdrückt. Für das Sichern numerischer Ergebnisse im Datenblatt ist ein separates Eingabefeld vorgesehen.

- **System-Optionen:** Optionen, die auf das ganze System angewendet werden.

- **Six-Sigma-Menü verwenden:** ein Six-Sigma-Menü mit den Überschriften, die der DMAIC-Anordnung entsprechen (Definieren, Messen, Analysieren, Verbessern, Regeln), wird angezeigt. Dieselben Prozeduren sind auch über das klassische Menü zugänglich, jedoch befinden sie sich hier unter anderen Stichworten.

- **Variablennamen sortieren:** gibt an, dass die Spaltennamen in Dateneingabe-Dialogboxen in alphabetischer Reihenfolge aufgelistet werden sollen. Andernfalls werden die Spaltennamen ihrer Reihenfolge im Datenblatt entsprechend angezeigt.

- **Vierstellige Jahreszahlen:** gibt an, dass Jahreszahlen mit vier anstatt zwei Stellen angezeigt werden sollen. Bei zweistelligen Jahreszahlen wie 1.2.05 wird angenommen, dass die Jahreszahl zwischen 1950 und 2049 liegt. Veränderungen dieser Option werden erst nach dem Neustart des Programms aktiviert.

- **Automatisch speichern aktiviert:** gibt an, dass die aktuellen *StatFolio*- und Datendateien im Hintergrund in dem festgelegten Intervall automatisch gesichert werden. Sollte eine Programmstörung auftreten, während diese Option aktiviert ist, wird Ihnen beim nächsten Programmstart angeboten, die betroffenen StatFolios und die Datenblätter wiederherzustellen.

- **Verknüpfungen für jeden Wert aktualisieren**: gibt an, dass alle Statistiken neu berechnet werden sollen, sobald ein Datenwert in einem der Datenblätter verändert wird. Ansonsten werden die Statistiken erst neu errechnet, wenn das Analyse-Fenster aktiviert, gedruckt oder publiziert oder wenn das StatFolio gesichert wird.

- **StatAdvisor**: hier können Standards für den StatAdvisor eingegeben werden.

    - **Hinzufügen zum Tabellen-Ergebnisfenster :** gibt an, dass die Ergebnisse des StatAdvisor automatisch am unteren Rand der Tabellen-Ergebnisfenster eingefügt werden. Die Ergebnisse des StatAdvisor sind zudem jederzeit über die entsprechende Schaltfläche auf der Symbolleiste verfügbar.

    - **Verweise hervorheben …:** gibt an, dass die Werte in den Tabellen-Ergebnisfenstern, auf die sich der StatAdvisor bezieht, in einer bestimmten Farbe hervorgehoben werden sollen.

- **Analyse-Überschrift:** gibt an, dass die Analyse-Überschrift am oberen Rand des Fensters *Analyse-Zusammenfassung* in blauer Schrift angezeigt werden soll.

- **StatFolios**: wenn das Kontrollkästchen *Start-Script deaktivieren* markiert ist, wird beim Laden eines StatFolio kein Script ausgeführt.

- **Verzeichnis für temporäre Dateien:** Ist hier ein Verzeichnis angegeben, dann werden StatFolios, Datendateien und andere Dateien zunächst in dieses Verzeichnis geschrieben und später an die endgültige Speicherstelle kopiert. Ist ein lokales Verzeichnis angegeben, kann dies die Speicherzeit im Vergleich mit der Zeit zum Speichern über verschiedene Netzwerke deutlich verkürzen.

Die Optionen in den weiteren Registern werden in dem PDF-Dokument *Preferences* ausführlich beschrieben.

9.2 Drucken

Die Druckausgabe-Optionen werden über zwei Auswahlmöglichkeiten im *Datei*-Menü geregelt:

1. *Druckeinrichtung*: ruft die Druckeroptionen Ihres Druckers auf. In dieser Dialogbox können normalerweise die Papiergröße und Hoch- bzw. Querformat eingestellt werden.

2. *Seite einrichten*: eine STATGRAPHICS Centurion XVI-spezifische Dialogbox, über welche die Seitenränder, Kopfzeilen des Drucks u. ä. eingestellt werden können. Die Optionen der Dialogbox wurden in Abschnitt 3.3 besprochen.

9.3 Diagramme

Wenn ein Ergebnisfenster mit einem Diagramm im Analyse-Fenster maximiert wird, wird damit die Schaltfläche *Diagramm-Optionen* auf der Analyse-Symbolleiste aktiviert. Mit dieser Schaltfläche wird eine Dialogbox mit mehreren Registern aufgerufen, über die das Erscheinungsbild des Diagramms bearbeitet werden kann (siehe Kapitel 4). Die Box enthält zudem das Register *Profil*, über das die eingestellten Diagramm-Eigenschaften in einem Benutzerprofil gesichert werden können sowie das Standardprofil für die Erstellung eines Diagramms modifiziert werden kann:

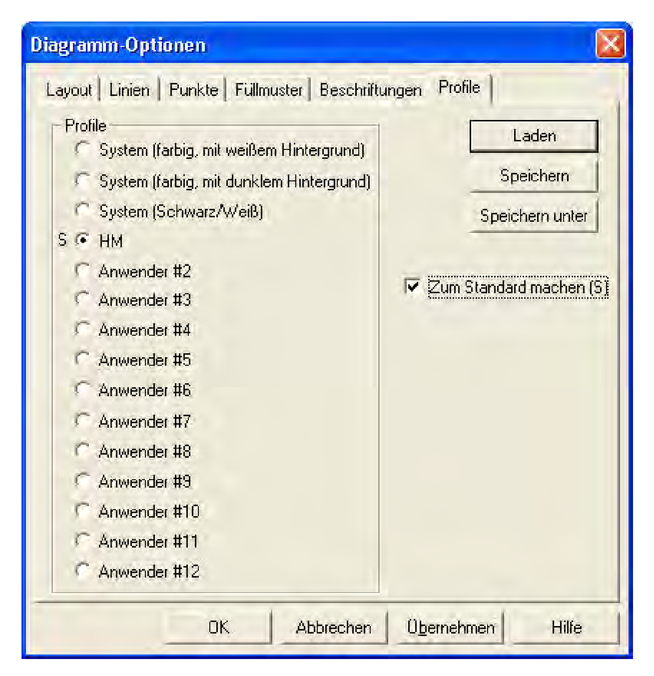


Abbildung 9.2: Profil-Register in der Dialogbox Diagramm-Optionen

Um die Standardwerte zu verändern, führen Sie folgende Schritte durch:

1. Bearbeiten Sie die Eigenschaften eines Diagramms in einem beliebigen Analyse-Fenster. Stellen Sie Farben, Schrifttypen und weitere Optionen so ein, wie sie auch auf künftige Diagramme angewendet werden sollen.

2. Wählen Sie auf der Analyse-Symbolleiste *Diagramm-Optionen* und klicken Sie das Register *Profil* an.

3. Markieren Sie das Kontrollkästchen *Zum Standard machen*.

4. Markieren Sie eins der 12 Anwenderprofile und wählen Sie die Schaltfläche *Speichern unter* (die Systemprofile können nur gelesen, nicht geschrieben werden).

5. Geben Sie dann einen Namen für das zu sichernde Profil ein:

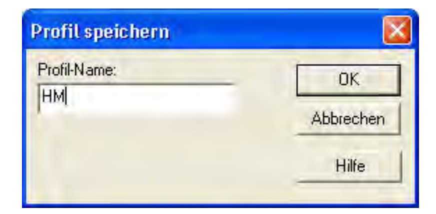


Abbildung 9.3: Dialogbox zum Sichern eines Profils

6. Wählen Sie *OK*, um die neu eingestellten Diagramm-Eigenschaften (Farben, Schriftarten, Punkt- und Linientypen, etc.) in einem neuen Profil zu sichern,

Das nächste Diagramm wird mit den in diesem Profil gesicherten Einstellungen erstellt.

Sie können auch andere gesicherte Profile auf ein neues Diagramm anwenden, indem Sie das Diagramm mit den Standardeinstellungen generieren und dann:

1. auf der Analyse-Symbolleiste *Diagramm-Optionen* wählen und das Register *Profil* anklicken.

2. eines der 15 Profile auswählen und die Schaltfläche *Laden* anklicken.

Das aktuelle Diagramm wird dann, den Eigenschaften des gewählten Profils entsprechend, aktualisiert.

Kapitel

10

Tutorial 1: Analysieren einer Stichprobe

Summenstatistiken, Häufigkeitsdiagramm, Box-Whisker-Plot, Konfidenzintervall und Hypothesen-Tests.

Eine häufige Fragestellung in der Statistik ist die nach der Analyse einer Stichprobe aus n Beobachtungen einer Grundgesamtheit. Betrachten wir z.B. die folgenden Körpertemperaturen von n = 130 Personen:

98,4	98,4	98,2	97,8	98	97,9	99	98,5	98,8	98
97,4	98,8	99,5	98	100,8	97,1	98	98,7	98,9	99
98,6	97,7	96,7	98,8	98,2	97,5	97,2	97,4	97,1	96,7
99,2	97,9	98,8	97,6	98,6	98,8	98,5	98,7	97,5	97,9
97,1	98,4	97,4	98,6	97,8	98,2	98	98	98,3	98,6
98,8	98,7	98,8	98,1	96,4	98,8	98,7	97,9	98,6	99,2
98,6	98	99,1	97,8	97,2	98,2	98,7	98,4	98,2	97,7
98,3	98,7	96,8	98	97,2	97,9	96,9	98,3	97,8	97
98,6	98,4	98,2	98	98	98,2	97,8	99	98,1	97,7
97,4	98,8	99,3	98,9	96,3	97,8	99,9	98,4	99,4	98,7
98,4	98,2	99,3	98,5	98,3	99	99,2	97,6	99,1	97,6
98,4	97,6	98,4	98	98,8	97,3	98,7	98,6	99,4	100
98,6	98,3	98,6	97,4	98,1	97,8	98,2	99	99,1	98,2

Diese Daten stammen aus dem Journal of Statistical Education Data Archive (www.amstat.org/publications/jse/jse_data_archive.html) und werden mit Erlaubnis verwendet.

Die Daten wurden in der Datei *bodytemp.sgd* gespeichert, die 130 Zeilen – eine Zeile pro an der Studie teilnehmender Person – enthält. Hier befinden sie sich in der Spalte *Temperature*.

Um eine einzelne Stichprobe aus einer Grundgesamtheit zusammenzufassen, bietet sich in STATGRAPHICS Centurion XVI besonders die Prozedur *Analyse einer Variablen* an. In dieser Prozedur werden die Daten in numerischer und grafischer Form zusammengefasst sowie Hypothesen über Mittelwert, Median und Standardabweichung der Grundgesamtheit getestet.

10.1 Ausführen der Prozedur *Analyse einer Variablen*

Um die Körpertemperatur-Daten auszuwerten, laden Sie die Datei *bodytemp.sgd* in ein Datenblatt. Hierfür führen Sie die folgenden Schritte aus:

1. Wählen Sie *Datei – Öffnen – Datenquelle öffnen* im Hauptmenü.
2. In der Dialogbox *Datenquelle öffnen* geben Sie an, dass Sie eine *STATGRAPHICS Datendatei* öffnen möchten.
3. In der Dialogbox *Datendatei öffnen*, markieren Sie die Datei *bodytemp.sgd* in der Liste der vorhandenen Dateien.

Nachdem Sie die Auswahl mit *OK* bestätigt haben, sollten die Daten in der folgenden Form in das Datenblatt geladen werden:

C:\Programme\Statgraphics\STATGRAPHICS Centurion XVI\Data\bod...

	Temperature degrees	Gender	Heart Rate beats per minute	Col_4	Col_5
1	98.4	Male	84		
2	98.4	Male	82		
3	98.2	Female	65		
4	97.8	Female	71		
5	98	Male	78		
6	97.9	Male	72		
7	99	Female	79		
8	98.5	Male	68		
9	98.8	Female	64		
10	98	Male	67		
11	97.4	Male	78		
12	98.8	Male	78		

bodytemp / B / C / D / E / F / G / H

Abbildung 10.1: Datenblatt mit den Körpertemperatur-Daten

Die Körpertemperatur-Daten, gemessen in Grad Fahrenheit, befinden sich in der linken Spalte.

Sie können die Prozedur *Analyse einer Variablen* an folgender Stelle im Hauptmenü aufrufen:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beschreiben – Numerische Daten – Analyse einer Variablen*.
2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Messbare Merkmale – Analyse einer Variablen*.

In der Dateneingabe-Dialogbox fügen Sie den Namen der auszuwertenden Spalte in das Eingabefeld *Daten* ein:



Abbildung 10.2: Dateneingabe-Dialogbox für die Analyse einer Variablen

Da alle 130 Zeilen analysiert werden sollen, lassen Sie das Eingabefeld *Bedingung* frei. Wählen Sie nun *OK*. Nachdem *OK* angeklickt wurde, erscheint die Dialogbox *Tabellen und Grafiken*. Hier können Sie alle verfügbaren Tabellen und Diagramme sehen und auswählen. Für unser Beispiel genügen jedoch die Standardeinstellungen.

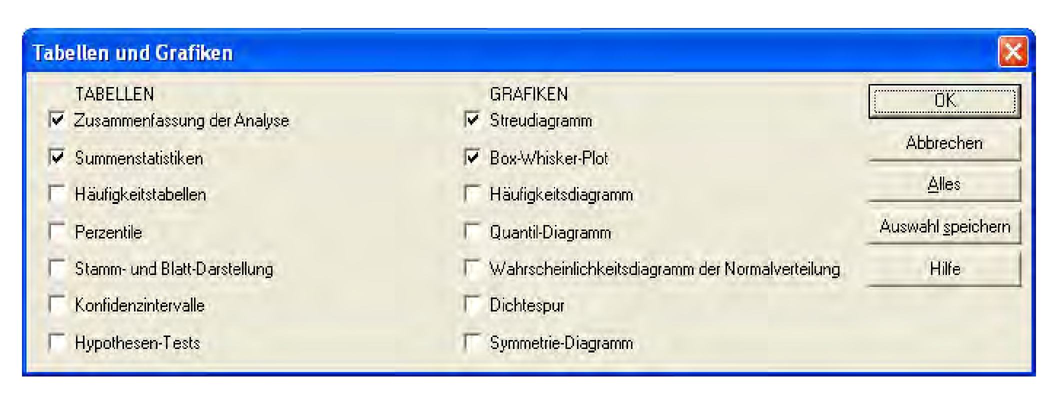


Abbildung 10.3: Dialogbox Tabellen und Grafiken

Das Programm erstellt daraufhin ein Analyse-Fenster, bestehend aus mehreren kleineren Ergebnisfenstern:

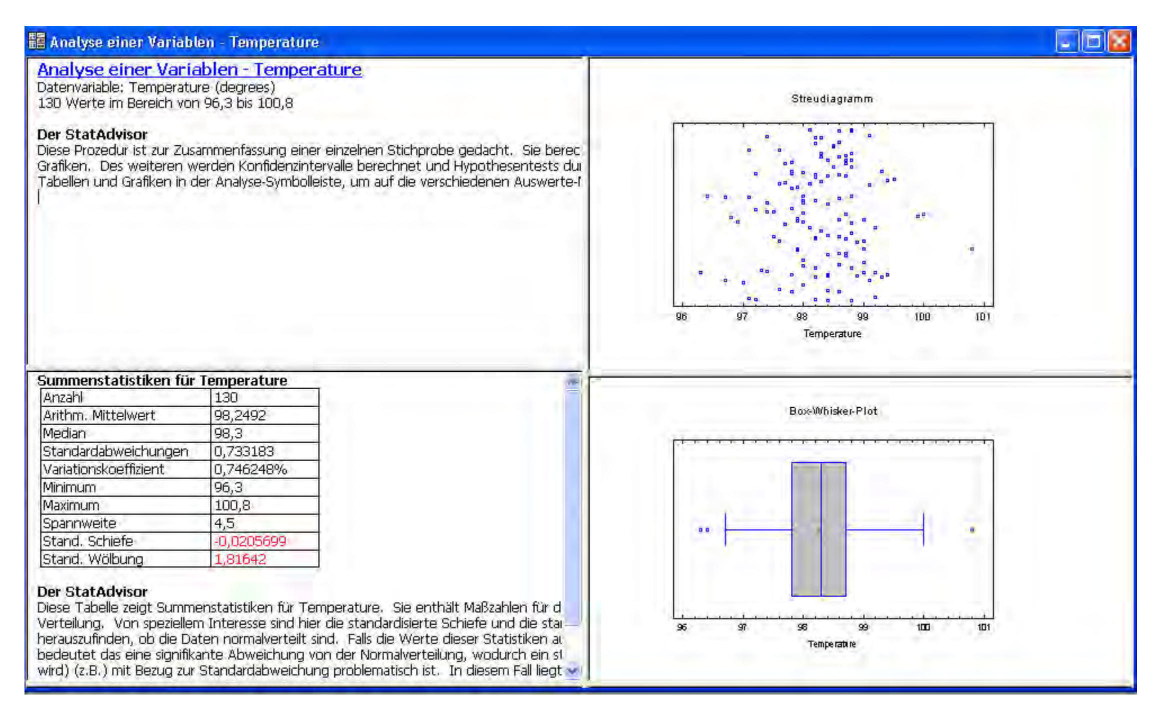


Abbildung 10.4: Analyse-Fenster der Analyse einer Variablen

Im linken oberen Ergebnisfenster befindet sich in der Regel die Zusammenfassung der Analyse. Dieser können Sie z. B. entnehmen, dass die Stichprobe $n = 130$ Werte umfasst, die sich in dem Bereich zwischen 96,3 und 100,8 Grad Fahrenheit befinden. Das rechte obere Ergebnisfenster enthält in der Regel ein Streudiagramm, bei dem die Datenpunkte in vertikaler Richtung zufällig angeordnet sind. Beachten Sie, dass die Punkte zwischen 98 und 99 Grad am dichtesten und nach den Seiten hin ausgedünnt sind. Dieses Verhalten ist typisch für Stichproben aus Grundgesamtheiten mit einem wohldefinierten zentralen Gipfel.
In den beiden unteren Ergebnisfenstern werden Summenstatistiken und ein Box-Whisker-Plot angezeigt, die in den folgenden Abschnitten näher beschrieben werden.

10.2 Summenstatistiken

Die Tabelle in diesem Ergebnisfenster zeigt verschiedene Summenstatistiken an. Wenn Sie weitere Statistiken benötigen, maximieren Sie dieses Ergebnisfenster (per Doppelklick in das Fenster) und wählen die Schaltfläche *Ergebnisfenster-Optionen* auf der Analyse-Symbolleiste:

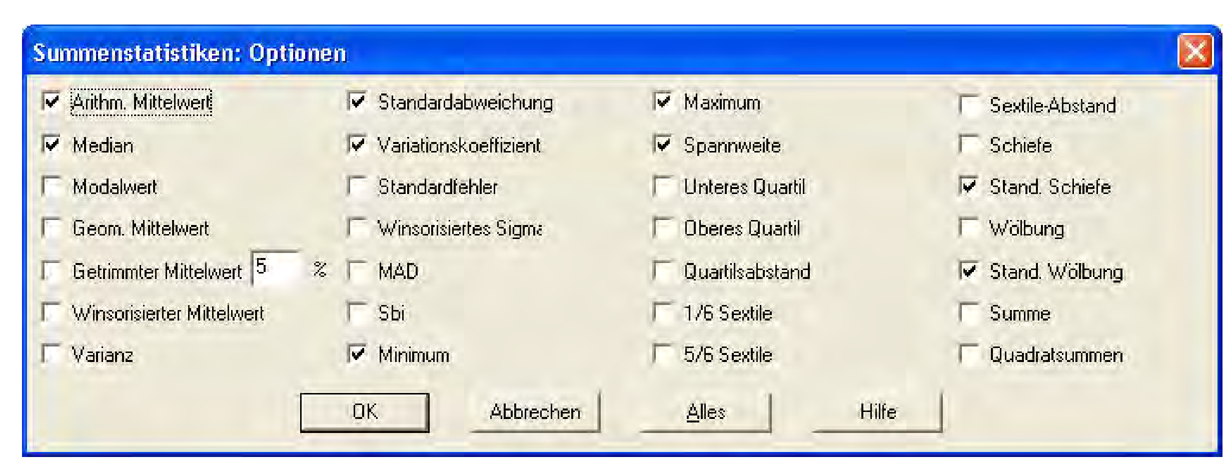


Abbildung 10.5: Dialogbox mit den Optionen für die Summenstatistiken

Wenn Sie hier noch die Quartile und den Quartilsabstand markieren, wird die Tabelle folgendermaßen ergänzt:

Summenstatistiken für Temperature

Anzahl	130
Arithm. Mittelwert	98,2492
Median	98,3
Standardabweichungen	0,733183
Variationskoeffizient	0,746248%
Minimum	96,3
Maximum	100,8
Spannweite	4,5
Unteres Quartil	97,8
Oberes Quartil	98,7
Quartilsabstand	0,9
Stand. Schiefe	-0,0205699
Stand. Wölbung	1,81642

Abbildung 10.6: Tabelle der Summenstatistiken

Bei Messdaten wird häufig angenommen, dass diese aus einer Normal- oder Gauss-Verteilung stammen, d.h. von einer glockenförmigen Verteilung. Daten aus einer Normalverteilung können mit Hilfe von zwei Statistiken vollständig beschrieben werden:

1. Stichproben-*Mittelwert* oder -*Durchschnitt* $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 98{,}25$, der das Zentrum der Verteilung schätzt.

2. Stichproben-*Standardabweichung* $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 0{,}733$, welche die Ausbreitung der Verteilung charakterisiert.

Bei einer Normalverteilung liegen ungefähr 68 % aller Werte innerhalb plus/minus einer Standardabweichung vom Mittelwert der Grundgesamtheit, ungefähr 95 % liegen innerhalb von plus/minus zwei Standardabweichungen und ungefähr 99,73 % der Werte liegen innerhalb von plus/minus drei Standardabweichungen.

Der Mittelwert und die Standardabweichung beschreiben eine Stichprobe nur dann vollständig, wenn sie einer Normalverteilung entstammt. Mit den Statistiken der standardisierten Schiefe und Wölbung kann diese Annahme überprüft werden. Diese Statistiken bewerten die Gestalt:

1. *Schiefe* misst die Symmetrie einer Verteilung bzw. das Fehlen dieser. Eine symmetrische Verteilung, wie z. B. die Normalverteilung, besitzt eine Schiefe von 0. Verteilungen, deren Werte eher *rechts vom* Gipfel als darunter liegen, besitzen Rechtsschiefe. Verteilungen, deren Werte eher *links vom* Gipfel liegen als darüber, weisen Linksschiefe auf.

2. *Wölbung* misst die Gestalt einer symmetrischen Verteilung. Eine Normal- oder Glockenkurve besitzt eine Wölbung von 0. Eine Verteilung, die einen spitzeren Gipfel als eine Normalverteilung aufweist, hat eine positive Wölbung. Eine Verteilung, die *flacher* als eine Normalverteilung ist, hat eine negative Wölbung.

Stammen die Daten aus einer Normalverteilung, sollten die Werte der standardisierten Schiefe und standardisierten Wölbung zwischen -2 und +2 liegen. Trifft dies zu, so scheint die Normalverteilung ein angemessenes Modell für die Daten zu sein.

Eine andere hilfreiche Zusammenfassung der Daten nach John Tukey bieten die folgenden fünf Zahlen :

Minimum (kleinster Datenwert) = 96,3
Unteres Quartil (25. Perzentil) = 97,8
Median (50. Perzentil) = 98,3

Oberes Quartil (75. Perzentil) = 98,7
Maximum (größter Datenwert) = 100,8

Diese fünf Angaben teilen die Stichprobe in vier Viertel ein, welche die Basis für das Box-Whisker-Plot liefern, das im nächsten Abschnitt beschrieben wird.

ANMERKUNG: Wenn Sie über *Ergebnisfenster-Optionen* weitere Summenstatistiken auswählen, so wird diese Auswahl nur für die vorliegende Analyse übernommen. Um Standard-Statistiken auch für künftige Analysen einzustellen, öffnen Sie über das Menü *Bearbeiten* die Dialogbox *Voreinstellungen*. Im Register *Statistiken* können Sie hier die standardmäßig auszuführenden Statistiken für die *Analyse einer Variablen* festlegen (und andere Prozeduren, die Summenstatistiken ausgeben).

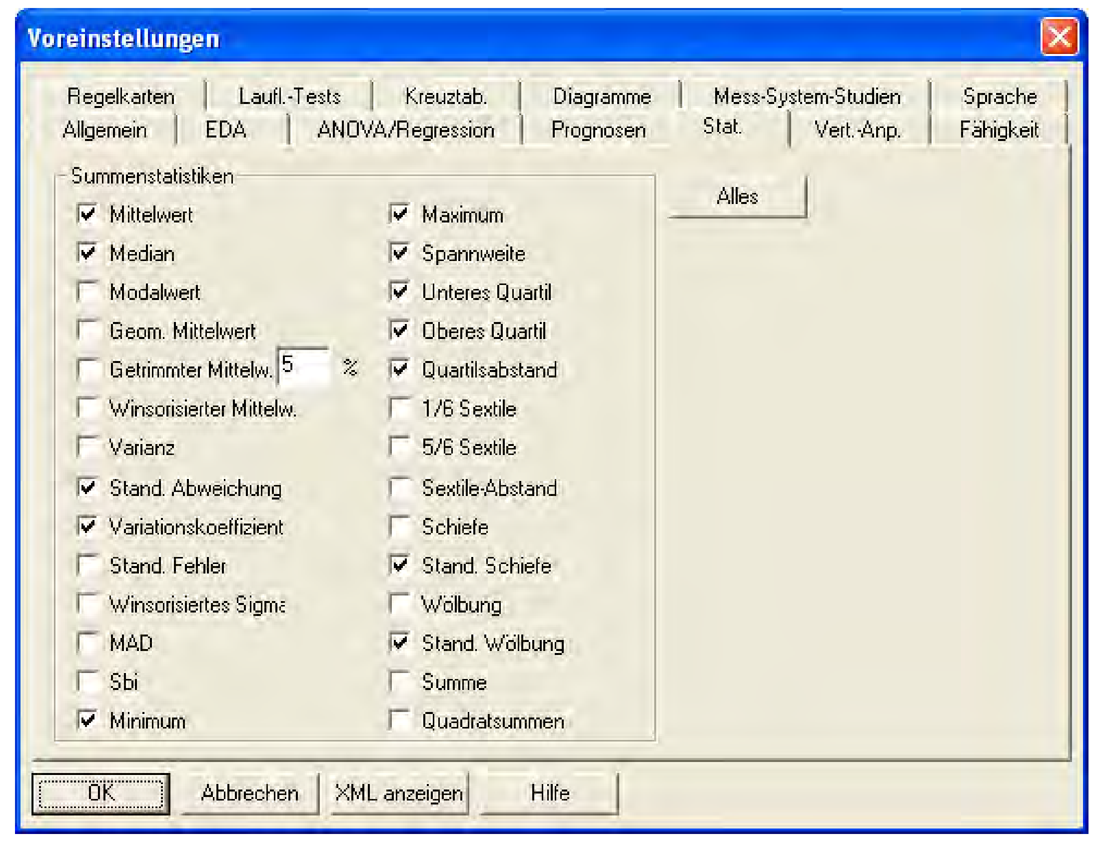


Abbildung 10.7: Voreinstellungen-Dialogbox zur Auswahl von Standard-Statistiken

10.3 Box-Whisker-Plot

Eine sehr schöne Grafik zur Zusammenfassung von Daten ist das Box-Whisker-Plot, das von John Tukey entwickelt wurde (siehe *Abb. 10.4*):

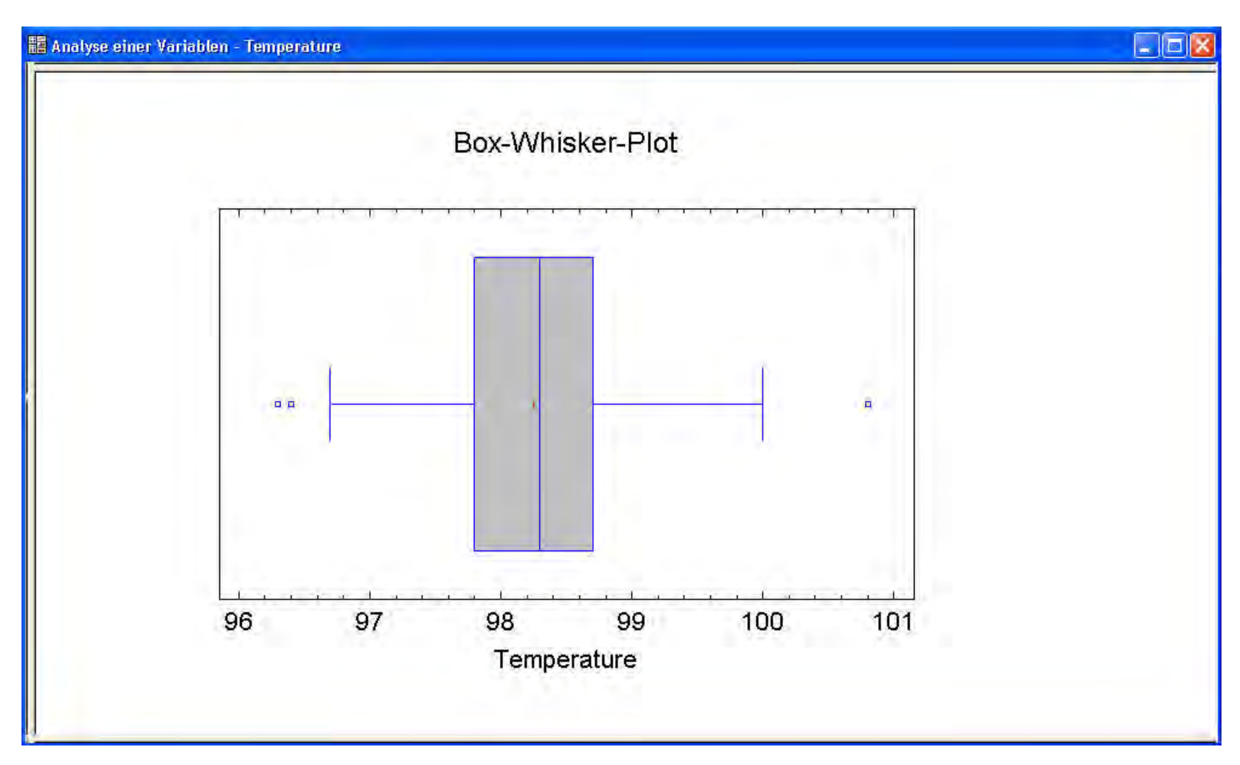


Abbildung 10.8: Box-Whisker-Plot der Körpertemperaturen

Ein Box-Whisker-Plot besteht aus den folgenden Elementen:

1. Eine Box, die sich vom unteren Quartil bis zum oberen Quartil erstreckt. Sie repräsentiert die mittleren 50 % der Datenwerte.

2. Eine vertikale Gerade an der Position des Stichproben-Medians, welche die Daten in zwei Hälften teilt. Kommen die Daten aus einer symmetrischen Verteilung, sollte sich diese Gerade in der Mitte der Box befinden.

3. Ein Plus-Symbol auf der Position des Mittelwerts der Stichprobe. Ein deutlicher Positionsunterschied zwischen dem Median und dem Mittelwert zeigt normalerweise an, dass Ausreißer (Datenwerte, die nicht aus der gleichen Population wie die restlichen

Daten stammen) oder eine schiefe Verteilung vorliegen. Im Falle einer schiefen Verteilung verschiebt sich der Mittelwert in Richtung der längeren Flanke.

4. Sogenannte Whisker oder Ausläufer (horizontale Linien), die sich von den Quartilen zu den Extremwerten der Stichprobe erstrecken, sofern es sich dabei nicht um "außerhalb liegende Punkte" (Ausreißer) handelt. Liegen einige Werte weit genug außerhalb der Box, um als "außerhalb liegende Punkte" bezeichnet werden zu können, dann reichen die Whisker bis zu den größten/kleinsten Werten, die nicht als "außerhalb liegend" betrachtet werden. STATGRAPHICS Centurion XVI ordnet solche ungewöhnlichen Punkte gemäß Tukey in zwei Kategorien ein:

    a. "Weit außerhalb liegende Punkte" (extreme Ausreißer) – Punkte, die sich mehr als den 3-fachen Quartilsabstand über- oder unterhalb der Grenzen der Box befinden. (Hinweis: Als Quartilsabstand (IQR) wird der Abstand zwischen den Quartilen bezeichnet, welcher gleichzeitig der Breite der Box entspricht.) Weit außerhalb liegende Punkte werden mit einem Punkt-Symbol (in der Regel einem kleinen Quadrat) und einem darüberliegenden Plus-Symbol dargestellt. Wenn die Daten aus einer Normalverteilung stammen, beträgt die Wahrscheinlichkeit, dass ein Punkt als "weit außerhalb liegend" betrachtet werden muss, in einer Stichprobe vergleichbarer Größe 1:300. Sofern die Stichprobe nicht mehrere Tausend Beobachtungen umfasst, weisen weit außerhalb liegende Punkte in der Regel auf echte Ausreißer (oder auf eine Nicht-Normalverteilung) hin.

    b. "Außerhalb liegende Punkte" (milde Ausreißer) – Punkte, die mehr als den 1,5-fache Quartilsabstand ober- oder unterhalb der Grenzen der Box liegen. Außerhalb liegende Punkte werden mit einem Punkt-Symbol, jedoch ohne darüberliegendes Plus-Symbol, dargestellt. Selbst wenn die Daten aus einer Normalverteilung stammen, beträgt die Wahrscheinlichkeit, dass ein oder zwei außerhalb liegende Punkte in einer Stichprobe aus n=100 Beobachtungen vorkommen, ungefähr 50 %. In diesem Fall deuten außerhalb liegende Punkte nicht auf echte Ausreißer hin. Dennoch sollten solche Punkte genauer untersucht werden.

Das Box-Whisker-Plot in *Abbildung 10.8* ist annähernd symmetrisch. Die Ausläufer haben ungefähr die gleich Länge und Mittelwert und Median der Stichprobe befinden sich in ähnlichem Abstand zur Mitte der Box. Es sind zwar drei außerhalb liegende Punkte dargestellt, doch gibt es keine besonders weit außerhalb liegenden Punkte. Wenn Sie den rechtsäußeren außerhalb liegenden Punkt mit der Maus anklicken, wird angezeigt, dass er sich in Zeile 15 der Datei befindet.

Über die Schaltfläche *Ergebnisfenster-Optionen* auf der Analyse-Symbolleiste können Sie dem Diagramm eine Median-Kerbe hinzufügen:

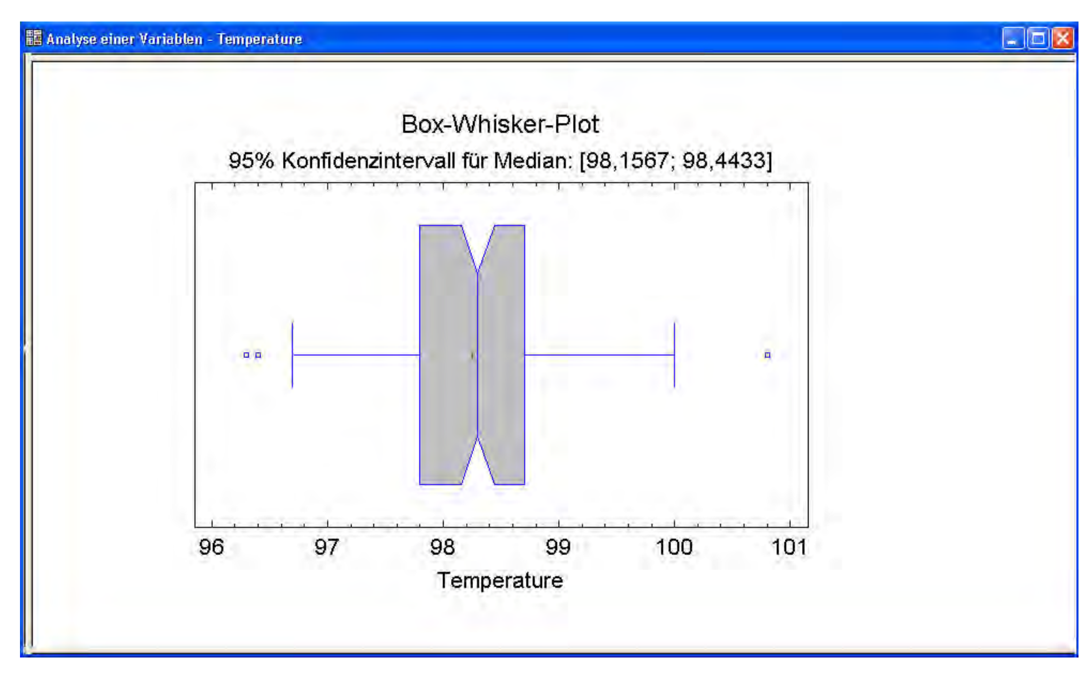


Abbildung 10.9: Box-Whisker-Plot mit einer 95 %-Median-Kerbe

In die Darstellung wird eine Kerbe eingefügt, die das ungefähre Konfidenzintervall für den Median der Grundgesamtheit abdeckt (zum standardmäßig eingestellten Konfidenzniveau 95 %). Dieses stellt den Fehlerbereich der Schätzung für die mittleren Temperatur der Grundgesamtheit mit Hilfe der Daten aus der Stichprobe dar. Im vorliegenden Fall beträgt der Stichprobenfehler ungefähr 0,15 Grad in beide Richtungen. Bei einer größeren Stichprobe würde die Fehlerspanne kleiner werden.

10.4 Ausreißer-Identifizierung

Bevor Sie weitere Statistiken schätzen, sollten Sie untersuchen, ob Zeile 15 einen echten Ausreißer darstellt und möglicherweise aus der Datenmenge ausgeschlossen werden sollte. STATGRAPHICS Centurion XVI enthält eine Prozedur, die einen formalen Test darüber ausführt, ob eine Beobachtung aus einer Normalverteilung stammen kann. Der Test befindet sich an folgender Stelle im Hauptmenü:

1. Im klassischen Menü *Beschreiben – Numerische Daten – Ausreißer-Identifizierung*.
2. Im Six Sigma-Menü *Analysieren – Messbare Merkmale – Ausreißer-Identifizierung*.

Wenn Sie *Temperature* in das *Daten*-Eingabefeld eingeben, wird zunächst Dialogbox *Optionen* und anschließend *Tabellen und Grafiken* aufgerufen. Nachdem Sie alle benötigten Optionen eingestellt haben, werden mehrere Tabellen erstellt, die in der linken Hälfte des Analyse-Fensters eingefügt werden. Besonders interessant ist die Tabelle, welche die 5 kleinsten und 5 größten Werte der Stichprobe auflistet:

Sortierte Werte

		Studentisierte Werte	*Studentisierte Werte*	*Modifiziert*
Zeile	*Wert*	*ohne Entfernen*	*mit Entfernen*	*MAD-z-Wert*
95	96,3	-2,65859	-2,74567	-2,698
55	96,4	-2,52219	-2,59723	-2,5631
23	96,7	-2,11302	-2,15912	-2,1584
30	96,7	-2,11302	-2,15912	-2,1584
73	96,8	-1,97663	-2,01521	-2,0235
...				
99	99,4	1,56955	1,59096	1,4839
13	99,5	1,70594	1,7323	1,6188
97	99,9	2,25151	2,30628	2,1584
120	100,0	2,3879	2,45231	2,2933
15	100,8	3,47903	3,67021	3,3725

Grubbs-Test (unter Annahme der Normalvert.)
Teststatistik = 3,47903
p-Wert = 0,0484379

Abbildung 10.10: Ausgewählte Ergebnisse der Prozedur zur Ausreißer-Identifizierung

Der ungewöhnlichste Wert befindet sich in Zeile 15 und wurde mit roter Schrift hervorgehoben. Er hat einen *studentisierten Wert (ohne Entfernen)* von 3,479. Studentisierte Werte werden berechnet aus:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Der Wert 3,479 bedeutet, dass sich die Beobachtung 3,479 Stichproben-Standarabweichungen oberhalb des Stichproben-Mittelwertes befindet (sofern diese Beobachtung in die Berechnung von $\bar{x}$ und s einbezogen wird). Die *studentisierten Werte (mit Entfernen)* zeigen an, wie viele Standardabweichungen sich jede Beobachtung vom Stichproben-Mittelwert entfernt befindet, wenn diese Beobachtung *nicht* in die Berechnungen mit einbezogen wird. Zeile 15 liegt 3,67 Standardabweichungen entfernt, wenn sie nicht in die Berechnungen einbezogen wird.

Beobachtungen, die einen Abstand von mehr als 3 Standardabweichungen zum Mittelwert besitzen, sind ungewöhnlich, es sei denn, der Stichprobenumfang n ist sehr groß ist oder es liegt

keine Normalverteilung vor. Ein formaler Test kann mit den folgenden Hypothesen durchgeführt werden:

> **Null-Hypothese**: Der extremste Wert stammt aus der gleichen Normalverteilung wie die anderen Beobachtungen.
>
> **Alternative Hypothese**: Der extremste Wert stammt nicht aus der gleichen Normalverteilung wie die anderen Beobachtungen.

Derartige Hypothesen werden häufig mit dem Grubbs-Test getestet, auch *Test auf extreme studentisierte Abweichungen* genannt. STATGRAPHICS Centurion XVI führt diesen Test durch und zeigt seinen *p-Wert* an. Im Allgemeinen quantifiziert der *p*-Wert die Wahrscheinlichkeit, eine von den in der Stichprobe beobachteten Werten in ungewöhnlicher oder stark ungewöhnlicher Weise abweichende Statistik zu erhalten, falls die Nullhypothese wahr ist. Wenn der *p*-Wert klein genug ist, kann die Null-Hypothese abgelehnt werden, da die Stichprobe ansonsten ein äußerst seltenes Ereignis gewesen wäre. "Klein genug" bedeutet in der Regel kleiner als 0,05, dieser Wert wird als "Signifikanzniveau" oder "Alpha-Risiko" des Tests bezeichnet. Wenn die Wahrscheinlichkeit für ein derartiges Stichprobenergebnis kleiner als 5 % ist – bei zutreffender Null-Hypothese – dann wird die Null-Hypothese verworfen.

Im vorliegenden Beispiel ist die Testgröße gleich dem größten absoluten *studentisierten Wert ohne Entfernen*: 3,479. Sie hat einen *p*-Wert von 0,0484. Da der *p*-Wert kleiner als 0,05 ist, würden wir die Null-Hypothese zurückweisen und dadurch schließen, dass Zeile 15 verglichen mit dem Rest der Daten einen Ausreißer darstellt. Sie können Zeile 15 entfernen, indem Sie die Schältfläche *Dateneingabe* auf der Analyse-Symbolleiste auswählen und anschließend im Eingabefeld *Bedingung* einen Ausdruck wie den folgenden eingeben:



Abbildung 10.11: Dialogbox zur Ausreißer-Identifizierung mit einer Bedingung zum Entfernen des Ausreißers

Da Zeile 15 die einzige Beobachtung enthält, die 100 Grad übersteigt, werden durch die Eingabe bei *Bedingung* nur noch die anderen *n* = 129 Zeilen ausgewertet. Die so modifizierten Ergebnisse der *Ausreißer-Identifizierung* sehen wie folgt aus:

Sortierte Werte

		Studentisierte Werte	*Studentisierte Werte*	*Modifiziert*
Zeile	*Wert*	*ohne Entfernen*	*mit Entfernen*	*MAD-z-Wert*
95	96,3	-2,75487	-2,85205	-2,698
55	96,4	-2,61209	-2,6956	-2,5631
23	96,7	-2,18375	-2,23455	-2,1584
30	96,7	-2,18375	-2,23455	-2,1584
73	96,8	-2,04097	-2,08332	-2,0235
...				
119	99,4	1,6713	1,69652	1,4839
99	99,4	1,6713	1,69652	1,4839
13	99,5	1,81408	1,84516	1,6188
97	99,9	2,3852	2,44992	2,1584
120	100,0	2,52798	2,60411	2,2933

Grubbs-Test (unter Annahme der Normalvert.)
Teststatistik = 2,75487
p-Wert = 0,676064

Abbildung 10.12: Ergebnisse der Ausreißer-Identifizierung nach Ausschluss der Zeile 15

Der extremste Wert unter den verbliebenen Beobachtungen befindet sich nun in Zeile 95. Da der *p*-Wert für Grubbs-Test deutlich über 0,05 liegt, scheinen alle der verbliebenen Beobachtungen aus der gleichen Grundgesamtheit zu kommen.

Idealerweise würde man nun zu der ursprünglichen Studie zurückkehren und versuchen, eine Ursache für den anormalen Wert in Zeile 15 zu ermitteln. Da dies hier nicht möglich ist, werden wir die Ergebnisse des Tests akzeptieren und Zeile 15 aus allen folgenden Berechnungen ausschließen. Wenn Sie die Dateneingabe in der Dialogbox für die *Analyse einer Variablen* modifizieren, wie in *Abbildung 10.11* gezeigt, werden folgende Summenstatistiken ausgegeben:

Summenstatistiken für Temperature

Anzahl	129
Arithm. Mittelwert	98,2295
Median	98,3
Standardabweichungen	0,70038
Variationskoeffizient	0,713004%
Minimum	96,3
Maximum	100,0
Spannweite	3,7
Unteres Quartil	97,8
Oberes Quartil	98,7
Quartilsabstand	0,9
Stand. Schiefe	-1,40217
Stand. Wölbung	0,257075

Abbildung 10.13: Summenstatistiken nach Ausschluss der Zeile 15

10.5 Häufigkeitsdiagramm

Eine weitere, häufig verwendete Grafik, die Messdaten aus einer Stichprobe veranschaulicht, ist das Häufigkeitsdiagramm. Wenn Sie zu der *Analyse einer Variablen* zurückkehren, können Sie ein solches Häufigkeitsdiagramm erzeugen, indem Sie die Schaltfläche *Tabellen und Grafiken* auf der Analyse-Symbolleiste auswählen und in der geöffneten Dialogbox das *Häufigkeitsdiagramm* aktivieren. Daraufhin wird ein standardmäßig ein Häufigkeitsdiagramm wie das folgende erstellt:

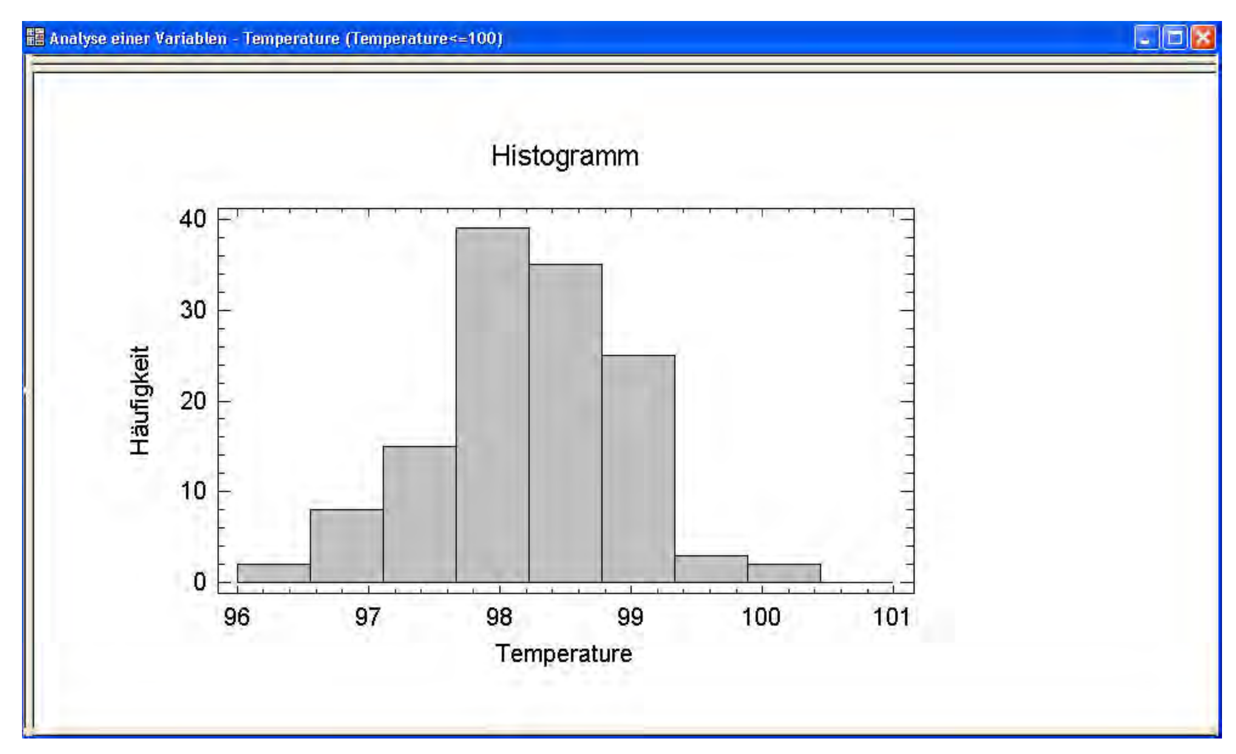


Abbildung 10.14: Häufigkeitsdiagramm mit Standard-Klasseneinteilung

Die Höhe jeder Säule des Diagramms gibt die Anzahl der Beobachtungen wieder, die in das *Temperature*-Intervall fallen, das von der entsprechenden Säule abgedeckt wird. Die Anzahl der Säulen und ihr Bereich wird standardmäßig, basierend auf dem Stichprobenumfang *n*, ausgegeben – je nachdem, welche Regel unter *Bearbeiten – Voreinstellungen – EDA* (Explorative Daten-Analyse) hierfür gewählt wurde:

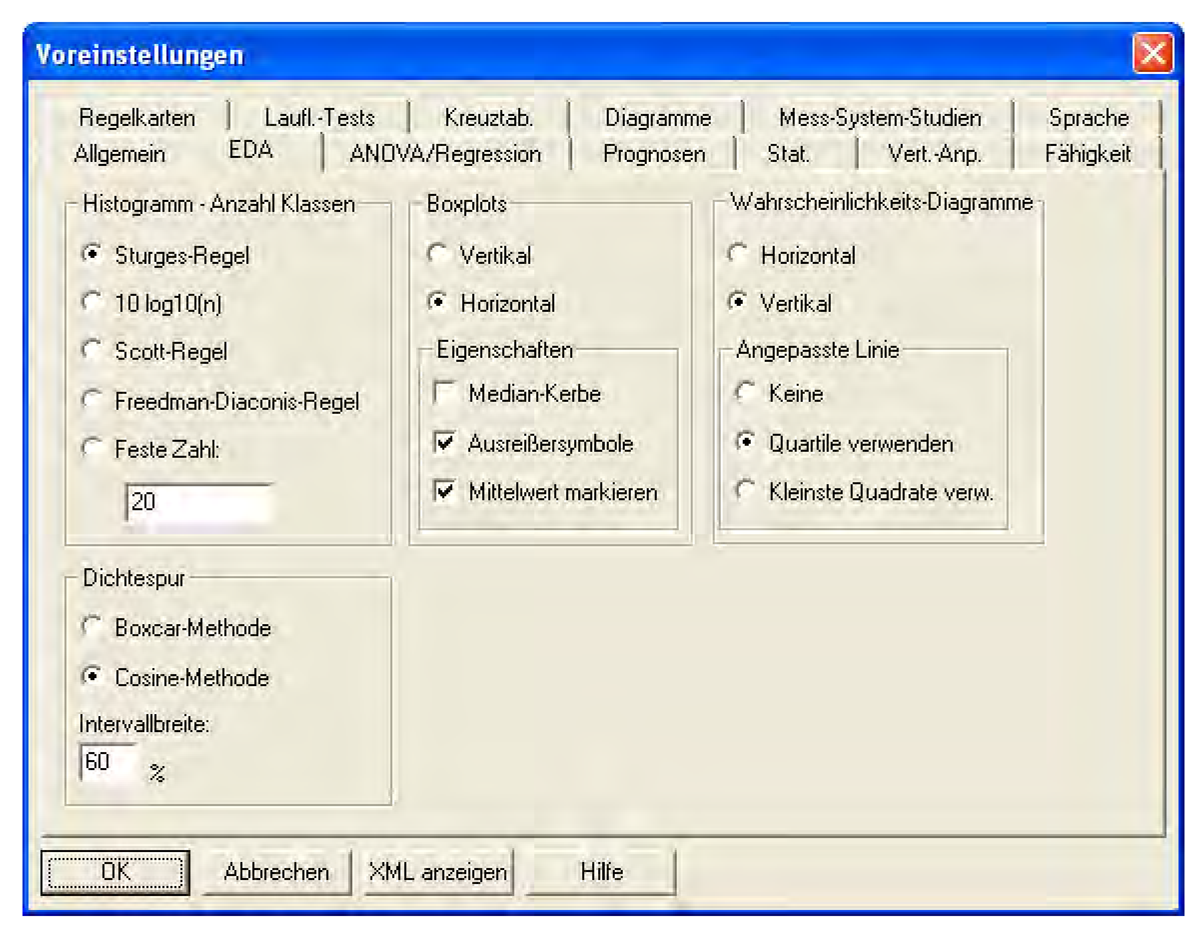


Abbildung 10.15: EDA-Register der Dialogbox Voreinstellungen

Mit der Sturges-Regel wird die Anzahl der Säulen auf die kleinste ganze Zahl eingestellt, die nicht kleiner als $(1+3{,}322\log_{10}(n))$ ist. Andere Regeln, wie z. B. die $10\log_{10}(n)$-Regel, tendieren dazu, mehr Säulen zu erzeugen. Dies kann von Vorteil sein, wenn Sie mit sehr großen Datenmengen arbeiten.

Sie können die Voreinstellungen zeitweise außer Kraft setzen, indem Sie das erzeugte Häufigkeitsdiagramm per Doppelklick maximieren und dann *Ergebnisfenster-Optionen* auf der Analyse-Symbolleiste auswählen:



Abbildung 10.16: Ergebnisfenster-Optionen-Dialogbox für das Häufigkeitsdiagramm

Beim Einstellen der Klassen sollte die Anzahl signifikanter Stellen in den Daten berücksichtigt werden. Die Körpertemperaturen wurden z.B. nur bis auf 0,1 Grad genau gemessen. Die Breite der Intervalle, die von den Säulen abgedeckt werden, sollte daher ein ganzzahliges Mehrfaches von 0,1 sein. Das folgende Diagramm enthält 25 Intervalle zwischen 96 und 100 Grad, wobei jede Säule ein Intervall von 0,2 Grad abdeckt:

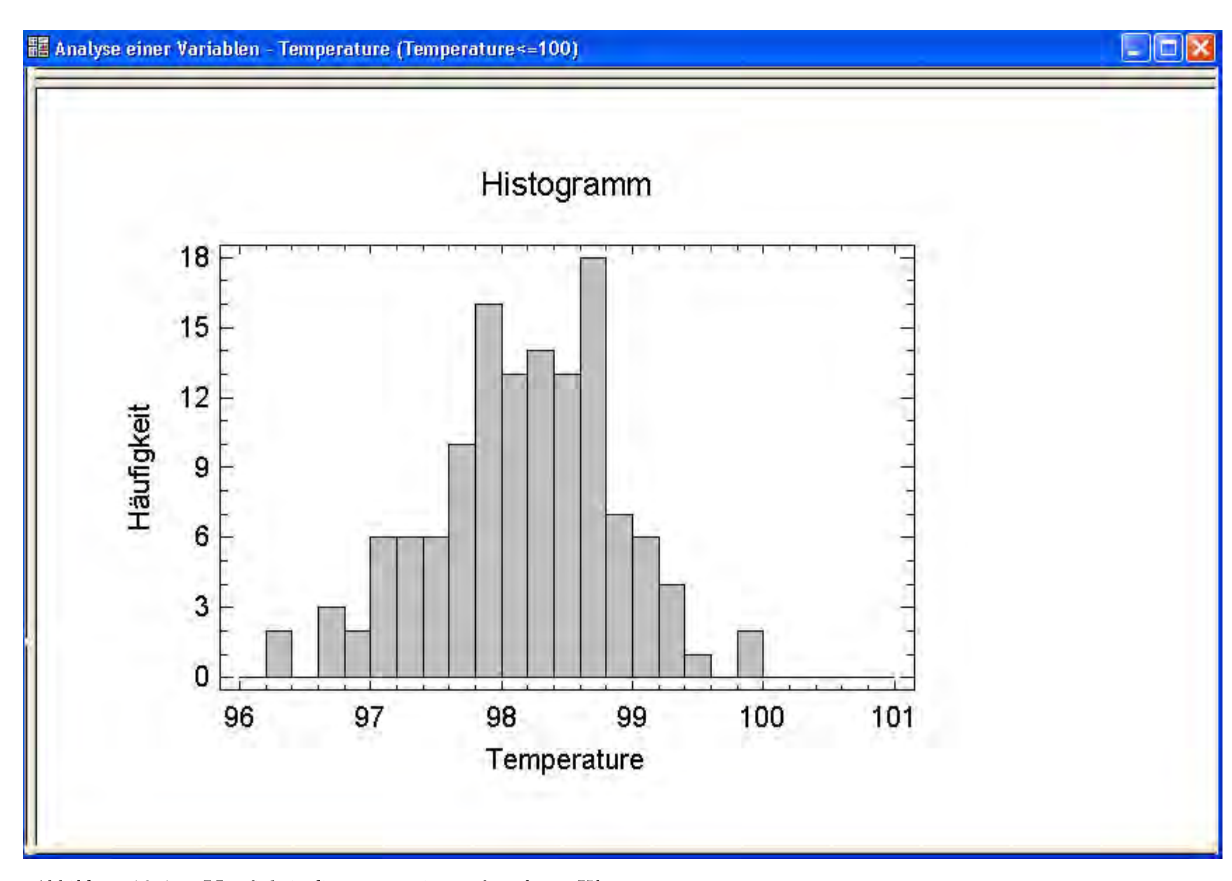


Abbildung 10.17: Häufigkeitsdiagramm mit neu festgelegten Klassen

Mit dieser größeren Anzahl an Klassen werden weitere Details sichtbar. Die allgemeine Gestalt der Verteilung ähnelt der glockenförmigen Normalverteilungskurve.

Die im Histogramm dargestellten Daten lassen sich auch in Tabellenform anzeigen, indem man die Schaltfläche *Tabellen und Grafiken* auf der Analyse-Symbolleiste auswählt und *Häufigkeitstabellen* aufruft:

Häufigkeitstabellen für Temperature

Klasse	Untere Grenze	Obere Grenze	Mittelpunkt	Häufigkeit	Relative Häufigkeit	Kumulative Häufigkeit	Kum. Rel. Häufigkeit
	bei oder unterhalb	96,0		0	0,0000	0	0,0000
1	96,0	96,2	96,1	0	0,0000	0	0,0000
2	96,2	96,4	96,3	2	0,0155	2	0,0155
3	96,4	96,6	96,5	0	0,0000	2	0,0155
4	96,6	96,8	96,7	3	0,0233	5	0,0388
5	96,8	97,0	96,9	2	0,0155	7	0,0543
6	97,0	97,2	97,1	6	0,0465	13	0,1008
7	97,2	97,4	97,3	6	0,0465	19	0,1473
8	97,4	97,6	97,5	6	0,0465	25	0,1938
9	97,6	97,8	97,7	10	0,0775	35	0,2713
10	97,8	98,0	97,9	16	0,1240	51	0,3953
11	98,0	98,2	98,1	13	0,1008	64	0,4961
12	98,2	98,4	98,3	14	0,1085	78	0,6047
13	98,4	98,6	98,5	13	0,1008	91	0,7054
14	98,6	98,8	98,7	18	0,1395	109	0,8450
15	98,8	99,0	98,9	7	0,0543	116	0,8992
16	99,0	99,2	99,1	6	0,0465	122	0,9457
17	99,2	99,4	99,3	4	0,0310	126	0,9767
18	99,4	99,6	99,5	1	0,0078	127	0,9845
19	99,6	99,8	99,7	0	0,0000	127	0,9845
20	99,8	100,0	99,9	2	0,0155	129	1,0000
21	100,0	100,2	100,1	0	0,0000	129	1,0000
22	100,2	100,4	100,3	0	0,0000	129	1,0000
23	100,4	100,6	100,5	0	0,0000	129	1,0000
24	100,6	100,8	100,7	0	0,0000	129	1,0000
25	100,8	101,0	100,9	0	0,0000	129	1,0000
	oberhalb	101,0		0	0,0000	129	1,0000

Mittelwert = 98,2295   Standardabweichungen = 0,70038

Abbildung 10.18. Häufigkeitstabelle

Eine Beobachtung wird als zu einem Intervall gehörig betrachtet, wenn sie größer als die untere Grenze und kleiner oder gleich der oberen Grenze des Intervalls ist.

Die rechte Spalte der Tabelle ist von besonderem Interesse, da sie die kumulative Wahrscheinlichkeit angibt, dass ein Wert in die gewählte Klasse oder eine vorhergehende Klasse fällt. So sind z. B. 89,92 % aller Datenwerte kleiner als oder gleich 99,0 Grad.

10.6 Quantil-Diagramm und Perzentile

Das Quantil-Diagramm, das Sie über die Schaltfläche *Grafiken* in der *Analyse einer Variablen* aufrufen können, sofern es nicht automatisch erzeugt wurde, stellt eine weitere Möglichkeit dar, kumulative Wahrscheinlichkeiten grafisch auszugeben:

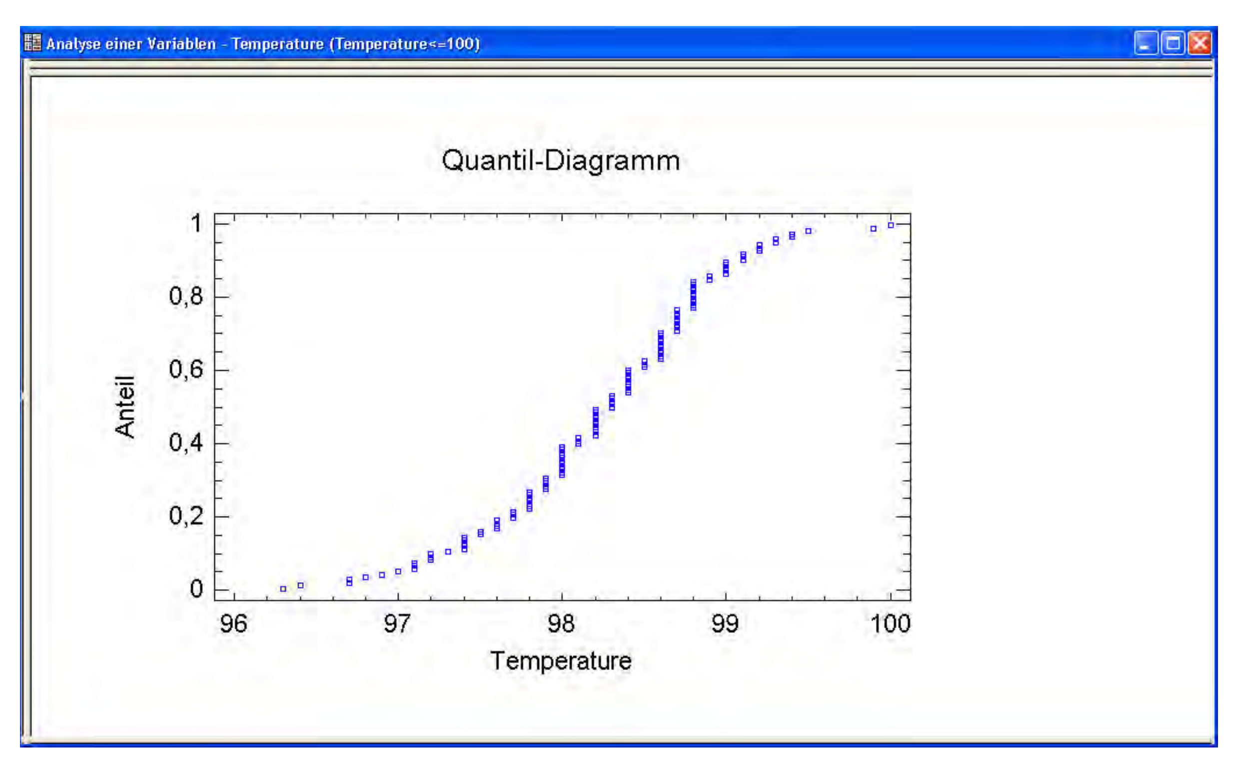


Abbildung 10.19: Quantil-Diagramm

Dieses Diagramm sortiert die Daten zunächst vom kleinsten zum größten Wert. Der *j*.-größte Datenwert wird dann auf Y = $(j+0,5)/n$ abgebildet. Dies schätzt den Anteil der Grundgesamtheit an oder unterhalb der beobachteten Temperatur. Wie die rechte Spalte in der Häufigkeitstabelle gibt die Kurve die kumulative Wahrscheinlichkeit wieder, dass eine Person eine Temperatur von weniger oder gleich der auf der horizontalen Achse abgebildeten Temperatur hat. Die vertikalen Sprünge im Diagramm sind dadurch zu erklären, dass die Temperatur-Daten nur auf 0,1 Grad genau gemessen wurden.

Abbildung 10.19 enthält zudem ein Fadenkreuz. Sie erzeugen so ein Fadenkreuz, indem Sie mit der rechten Maustaste in die Grafik klicken und im Kontextmenü *Lokalisieren* auswählen. Mit der Maus können Sie das Fadenkreuz an jede gewünschte Position bewegen. Die Position des Fadenkreuzes wird mit den Zahlen an den Fäden des Fadenkreuzes angegeben. Im obigen

Diagramm wurde das Fadenkreuz verwendet, um den Median (50. Perzentil) zu lokalisieren. Das 50. Perzentil ist der Wert von *temperature*, an dem der auf der vertikalen Achse angegebene Anteil 0,5 entspricht.

Sie können eine Tabelle der Perzentile erstellen, indem Sie über die Schaltfläche *Tabellen und Grafiken* das Kontrollkästchen *Perzentile* markieren:

Perzentile für Temperature

	Perzentile	*Untere Grenze*	*Obere Grenze*
1,0%	96,4	96,34	96,811
5,0%	97,0	96,8727	97,2473
10,0%	97,2	97,1538	97,4829
25,0%	97,8	97,6152	97,8846
50,0%	98,3	98,1082	98,3508
75,0%	98,7	98,5743	98,8437
90,0%	99,1	98,9761	99,3051
95,0%	99,3	99,2116	99,5862
99,0%	99,9	99,6479	100,119

Die Ausgabe enthält 95,0%-Konfidenzgrenzen auf Basis der Normalverteilung.

Abbildung 10.20: Perzentil-Tabelle

Das *p*.-Perzentil schätzt den Temperatur-Wert, unterhalb dem *p* % der Grundgesamtheit liegen. Mit *Ergebnisfenster-Optionen* wurden 95 %-Konfidenzgrenzen zu den Perzentilen hinzugefügt, basierend auf der Annahme, dass die Stichprobe aus einer Normalverteilung stammt.

Das 90%-Perzentil ist z. B. der Temperatur-Wert, der nur von 10 % der Personen aus der Grundgesamtheit überschritten wird. Der Schätzwert dieses Perzentils auf Basis der Stichprobe ist 99,1 Grad. Da aber der Stichprobenumfang sehr gering ist, könnte sich das wahre 90%-Perzentil mit 95%iger Wahrscheinlichkeit irgendwo zwischen 98,98 und 99,31 Grad befinden.

10.7 Konfidenzintervalle

Nachdem der Ausreißer aus der Stichprobe entfernt wurde, können Sie nun fortfahren, die endgültigen Schätzwerte für die Parameter der Verteilung, aus der die Daten stammen, zu bestimmen. Wählen Sie *Konfidenzintervalle* in der Dialogbox *Tabellen und Grafiken,* um Folgendes anzuzeigen:

Konfidenzintervalle für Temperature
95,0% Konfidenzintervall für Mittelwert: 98,2295 +/- 0,122015 [98,1074; 98,3515]
95,0% Konfidenzintervall für Standardabweichung: [0,624081; 0,798114]

Abbildung 10.21: 95 % Konfidenzintervalle für den Mittelwert und die Standardabweichung

Die Konfidenzintervalle beschränken den potenziellen Fehler bei der Schätzung des Mittelwerts und der Standardabweichung der Grundgesamtheit. Für die *n*=129 Beobachtungen können Sie

mit 95 %iger Wahrscheinlichkeit sagen, dass die mittlere Temperatur der Grundgesamtheit irgendwo zwischen 98,11 und 98,35 Grad liegt. Ebenso liegt die Standardabweichung der Grundgesamtheit irgendwo zwischen 0,624 und 0,798 Grad.

Mit *Ergebnisfenster-Optionen* können Sie weitere Konfidenintervalle unter Verwendung der Bootstrap-Methode anfordern:

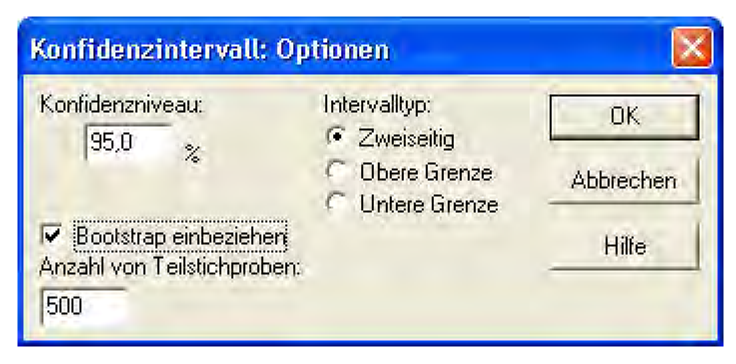


Abbildung 10.22: Konfidenzintervalle-Optionen

Bootstrap-Intervalle verlassen sich, anders als die Intervalle in *Abbildung 10.21*, nicht auf die Annahme, dass die Grundgesamtheit einer Normalverteilung folgt. Stattdessen werden zufällige Stichproben mit *n*=129 Beobachtungen aus den Daten gezogen, wobei gleiche Beobachtungen mehr als einmal gewählt werden können. Dies wird 500 Mal wiederholt, Statistiken werden berechnet und die mittleren 95 % der Ergebnisse werden verwendet, um die Konfidenzintervalle zu berechnen. Die folgende Tabelle zeigt Bootstrap-Intervalle für den Mittelwert, die Standardabweichung und den Median der Grundgesamtheit:

Konfidenzintervalle für Temperature
95,0% Konfidenzintervall für Mittelwert: 98,2295 +/- 0,122015 [98,1074; 98,3515]
95,0% Konfidenzintervall für Standardabweichung: [0,624081; 0,798114]

Bootstrap-Intervalle
Mittelwert: [98,1109; 98,3543]
Standardabweichung: [0,606691; 0,784368]
Median: [98,1; 98,4]

Abbildung 10.23: 95%-Bootstrap-Konfidenzintervalle

ANMERKUNG: Es ist möglich, dass Ihre Ergebnisse leicht von den obigen Ergebnissen abweichen.

Die Bootstrap-Intervalle sind den vorigen Intervallen, die mit der Student-t-Verteilung und der Chi-Quadrat-Verteilung berechnet wurden, sehr ähnlich. Diese Ähnlichkeit ist nicht unerwartet, weil die Daten keine signifikanten Schiefe und Wölbung besitzen.

10.8 Hypothesen-Tests

Es können auch formale Hypothesen-Tests durchgeführt werden. Z.B. wird häufig behauptet, dass die übliche menschliche Körpertemperatur 98,6 Grad Fahrenheit beträgt. Um zu testen, ob die vorliegenden Daten aus einer Normalverteilung mit diesem Mittelwert kommen, kann ein Hypothesen-Test für:

Null-Hypothese: $\mu = 98,6$ Grad

Alternative Hypothese: $\mu \neq 98,6$ Grad

formuliert werden. Um den Test innerhalb der *Analyse einer Variablen* auszuführen, wählen Sie *Hypothesen-Tests* aus der Liste in *Tabellen und Grafiken*. Bevor Sie die Ergebnisse untersuchen, wählen Sie *Ergebnisfenster-Optionen* und legen Sie die Eigenschaften für den gewünschten Test fest:



Abbildung 10.24: Ergebnisfenster-Optionen für Hypothesen-Tests

Der Wert, den Sie bei *Mittelwert/Median* eingeben, repräsentiert die Null-Hypothese. Im Gruppenfeld *Alt. Hypothese* wählen Sie die alternative Hypothese:

1. *Ungleich*: $\mu \neq 98.6$
2. *Kleiner als*: $\mu < 98.6$
3. *Größer als*: $\mu > 98.6$

Obwohl die Stichprobe eine geringere mittlere Temperatur vermuten ließe, wurde hier eine zweiseitige Alternative gewählt. Die Erstellung eines einseitigen Tests mit der alternativen Hypothese $\mu < 98.6$ Grad würde an dieser Stelle als "data snooping (Datenschnüffeln)" betrachtet werden, da wir die Hypothese formuliert würden, nachdem wir die Daten bereits gesichtet hätten.

Die Ergebnisse des Tests sehen wie folgt aus:

Hypothesen-Tests für Temperature
Stichpr.-Mittelwert = 98,2492
Stichpr.-Median = 98,3
Stichpr.-Standardabweichung = 0,733183

<u>t-Test</u>
Null-Hypothese: Mittelwert = 98,6
Alternative: ungleich

Berechnete t-Statistik = -5,45482
p-Wert = [illegible]
Nullhypothese ablehnen für alpha = 0,05.

<u>Vorzeichen-Rang-Test</u>
Null-Hypothese: Median = 98,6
Alternative: ungleich

Mittlerer Rang der Werte unterhalb des hypothetischen Medians: 67,7222
Mittlerer Rang der Werte oberhalb des hypothetischen Medians: 45,5

Teststatistik für große Stichproben = 4,86 (mit angewandter Korrektur für Bindungen)
p-Wert = [illegible]
Nullhypothese ablehnen für alpha = 0,05.

Abbildung 10.25: Ergebnisse des Hypothesen-Tests

Die Ergebnisse von zwei Tests werden angezeigt:

1. Der Standard-*t*-Test, der davon ausgeht, dass die Daten aus einer Normalverteilung kommen (der Test ist nicht sehr empfindlich gegenüber Abweichungen von dieser Annahme).

2. Der nicht-parametrische Vorzeichen-Rang-Test, basierend auf den Rangzahlen des Abstands jeder Beobachtung vom hypothetischen Median. Dieser Test setzt keine Normalverteilung voraus und ist weniger empfindlich gegenüber Ausreißern als der *t*-Test.

In beiden Fällen ist der *p*-Wert weit unter 0,05, was die Hypothese, dass die Stichprobe aus einer Grundgesamtheit mit dem Mittelwert 98,6 Grad stammt, sicher widerlegt.

> ANMERKUNG: E-8 nach einer Zahl bedeutet, dass die angegebene Zahl mit 10^{-8} multipliziert wird. Der *p*-Wert 1,81264E-8 entspricht demnach dem Wert 0,0000000181264.

Bitte beachten Sie, dass das Konfidenzintervall für den Mittelwert, angegeben in Abschnitt 10.8, den Wert 98,6 *nicht* mit einschloss. Werte, die sich nicht innerhalb des Konfidenzintervalls befinden, würden von dem hier besprochenen *t*-Test abgelehnt werden.

10.9 Toleranzgrenzen

Für die Körpertemperatur-Daten ist noch eine weitere Analyse interessant. Diese berechnet Toleranzgrenzen, das sind Grenzen, innerhalb derer für ein vorgegebenes Konfidenzniveau schätzungsweise ein bestimmter Prozentanteil der Grundgesamtheit liegt. Toleranzgrenzen werden über das Hauptmenü an folger Stelle aufgerufen:

1. Im klassischen Menü *Beschreiben –Numerische Daten – Statistische Toleranzgrenzen*
2. Im Six Sigma-Menü *Analysieren – Messbare Merkmale – Fähigkeitsanalyse – Statistische Toleranzgrenzen*

Die Prozedur startet mit der Anzeige einer Dialogbox, in der Sie den Stichprobenumfang *n* sowie den Mittelwert und die Standardabweichung der Probe eingeben. Wenn Sie die Ergebnisse aus *Abbildung 10.13* verwenden, sollte Ihre Eingabe folgendermaßen aussehen:

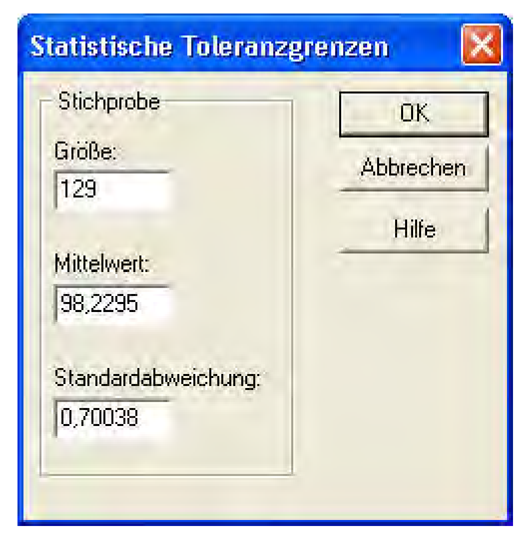
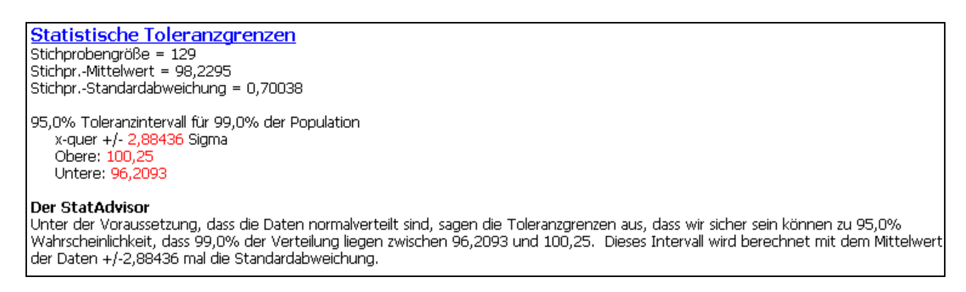


Abbildung 10.26: Dialogbox für Statistische Toleranzgrenzen

Nach dem Bestätigen mit *OK*, erscheint das *Optionen*-Menü und anschließend die Dialogbox *Tabellen und Grafiken*. Die Ergebnisse hieraus sehen wie folgt aus:



Statistische Toleranzgrenzen
Stichprobengröße = 129
Stichpr.-Mittelwert = 98,2295
Stichpr.-Standardabweichung = 0,70038

95,0% Toleranzintervall für 99,0% der Population
x-quer +/- 2,88436 Sigma
Obere: 100,25
Untere: 96,2093

Der StatAdvisor
Unter der Voraussetzung, dass die Daten normalverteilt sind, sagen die Toleranzgrenzen aus, dass wir sicher sein können zu 95,0% Wahrscheinlichkeit, dass 99,0% der Verteilung liegen zwischen 96,2093 und 100,25. Dieses Intervall wird berechnet mit dem Mittelwert der Daten +/-2,88436 mal die Standardabweichung.

Abbildung 10.27: Zusammenfassung der Analyse für Statistische Toleranzgrenzen

Die Interpretation des StatAdvisor fasst die Ergebnisse kurz und bündig zusammen. Das Konfidenzniveau sowie der Prozentanteil der Grundgesamtheit können über *Ergebnisfenster-Optionen* modifiziert werden.

Die Prozedur *Statistische Toleranzgrenzen* gibt zudem ein *Toleranzgrenzen-Diagramm* aus, das die Toleranzgrenzen in grafischer Form veranschaulicht:

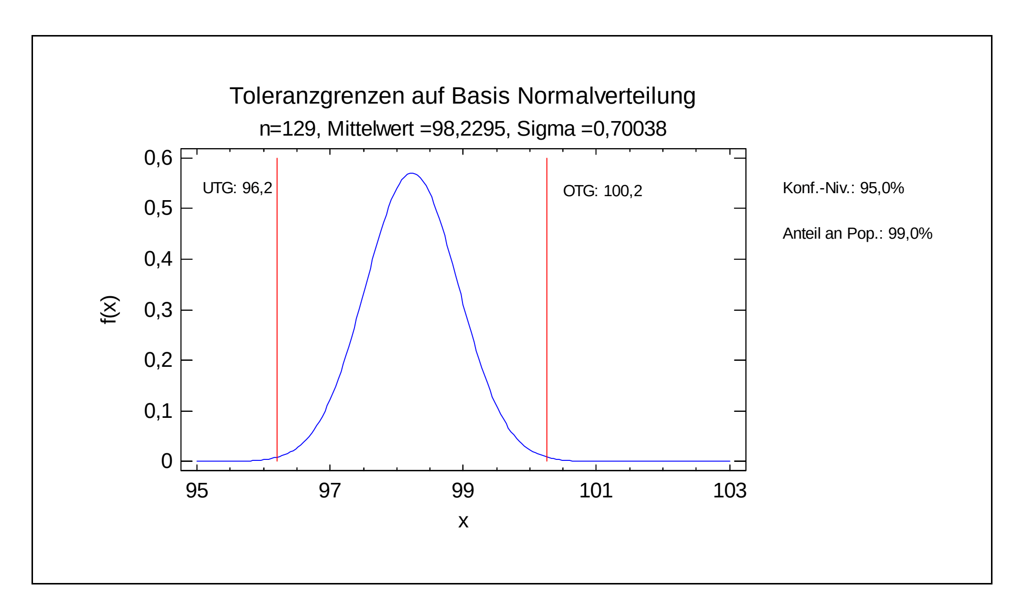


Abbildung 10.28: Toleranzgrenzen-Diagramm

Es ist unwahrscheinlich, dass sich mehr als eine von 100 Körpertemperaturen außerhalb der berechneten Grenzen befindet.

Kapitel

11

Tutorial 2: Vergleichen zweier Stichproben

Grafische Vergleiche und Hypothesen-Tests.

Häufig bestehen Daten, die analysiert werden sollen, aus zwei Stichproben, die möglicherweise aus verschiedenen Grundgesamtheiten stammen. In solchen Fällen ist es sinnvoll:

1. Die Daten so darzustellen, dass visuelle Vergleiche möglich sind.

2. Hypothesen über statistisch signifikante Unterschiede zwischen den Stichproben zu testen.

In Tutorial 1 im vorhergehenden Kapitel wurde eine Datenmenge mit Körpertemperaturen von 130 Probanden analysiert. Unter diesen 130 Probanden waren 65 weiblich und 65 männlich. Wir wollen nun in diesem Tutorial die Daten der Frauen mit denen der Männer vergleichen.

Um die Körpertemperaturen zu analysieren, öffnen Sie die Datendatei *bodytemp.sgd* unter Verwendung der Menüoption *Datenquelle öffnen* unter *Datei – Öffnen*.

11.1 Ausführen der Prozedur Vergleich zweier Stichproben

Die wichtigste Prozedur für den Vergleich von Daten aus zwei Stichproben ist *Vergleich zweier Stichproben*, die man im Hauptmenü folgendermaßen erhält:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Vergleichen – Zwei Stichproben – Unabhängige Stichproben.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Messbare Merkmale – Vergleich von zwei Stichproben – Unabhängige Stichproben.*

Die Dateneingabe-Dialogbox für diese Prozedur sieht wie folgt aus:

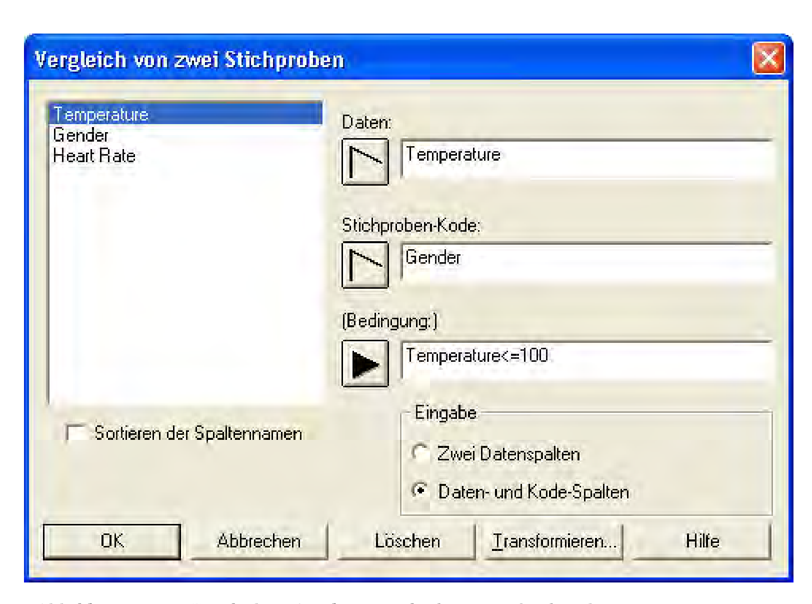


Abbildung 11.1: Dialogbox für den Vergleich zweier Stichproben

Die *Eingabe*-Box gibt an, in welcher Weise die Daten für die beiden Stichproben eingegeben werden sollen:

1. *Zwei Datenspalten* – die Daten für jede Stichprobe befinden sich in verschiedenen Spalten.
2. *Daten- und Kode-Spalte* – die Daten für beide Stichproben befinden sich in derselben Spalte, eine zweite Spalte enthält die Kodes zur Unterscheidung der beiden Stichproben.

Die Datei *bodytemp.sgd* besitzt letztere Struktur, mit allen n = 130 Beobachtungen in einer Spalte namens *Temperature* sowie einer zweiten Spalte namens *Gender,* welche die beschreibenden Werte „Female“ oder „Male“ enthält. Im Feld *Bedingung* wurde eine Eintragung vorgenommen, um nur diejenigen Zeilen zu verwenden, für die *Temperature* kleiner oder gleich 100 ist. Damit wird Zeile 15, die bereits in Kapitel 10 als Ausreißer erkannt wurde, aus der Analyse ausgeschlossen.

Nachdem die Dialogbox *Tabellen und Grafiken* angezeigt wurde, öffnet sich ein neues Analyse-Fenster. Es enthält vier Ergebnisfenster mit der Zusammenfassung der Analyse, den Summenstatistiken, einem Doppel-Histogramm und einem zweifachen Box-Whisker-Plot:

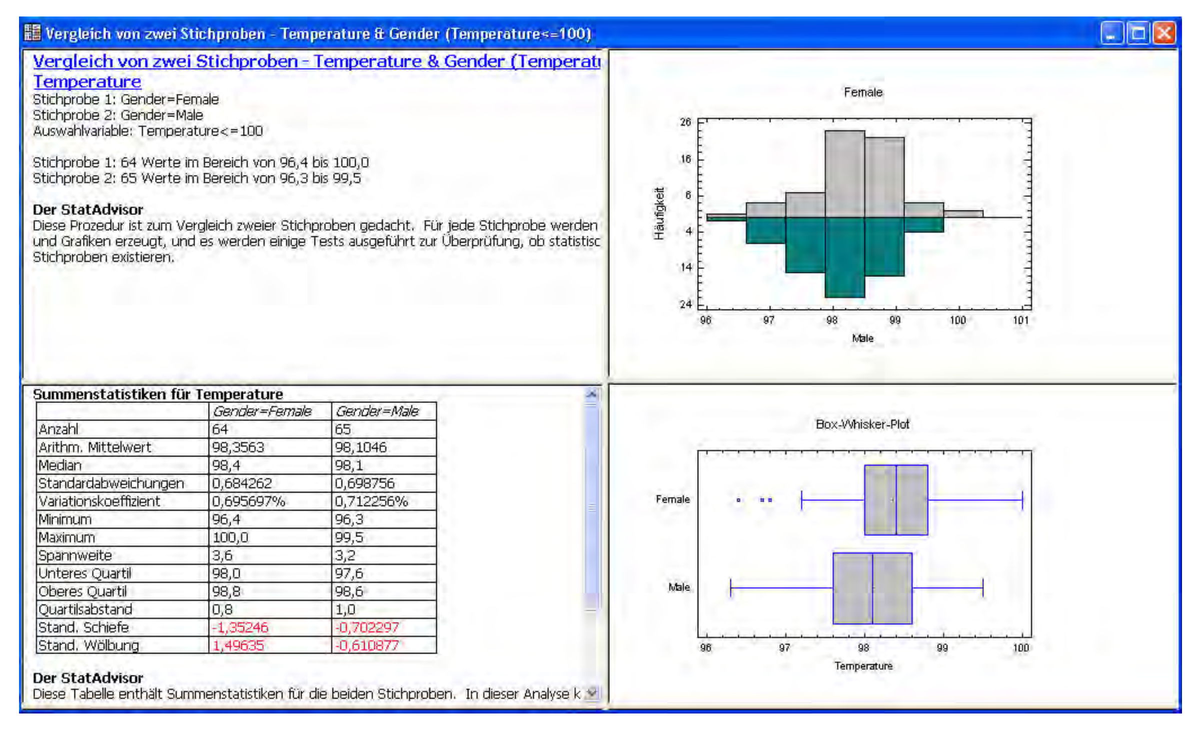


Abbildung 11.2: Analyse-Fenster für den Vergleich zweier Stichproben

Nach Entfernung des Ausreißers gibt es n_1 = 64 Beobachtungen für weibliche Probanden im Bereich von 96,4 bis 100,0 Grad und n_2 = 65 Beobachtungen für männliche im Bereich von 96,3 bis 99,5 Grad.

11.2 Summenstatistiken

Die *Summenstatistik*-Tabelle enthält statistische Maßzahlen für jede Stichprobe:

Summenstatistiken für Temperature		
	Gender=Female	*Gender=Male*
Anzahl	64	65
Arithm. Mittelwert	98,3563	98,1046
Median	98,4	98,1
Standardabweichungen	0,684262	0,698756
Variationskoeffizient	0,695697%	0,712256%
Minimum	96,4	96,3
Maximum	100,0	99,5
Spannweite	3,6	3,2
Unteres Quartil	98,0	97,6
Oberes Quartil	98,8	98,6
Quartilsabstand	0,8	1,0
Stand. Schiefe	-1,35246	-0,702297
Stand. Wölbung	1,49635	-0,610877

Abbildung 11.3: Summenstatistiken je Stichprobe

Mehrere Dinge sind von besonderem Interesse:

1. Die Durchschnittstemperatur der Frauen ist etwa 0,25 Grad höher als die der Männer. Die Differenz zwischen den Medianen beträgt 0,30 Grad.

2. Die Standardabweichung der Frauen ist leicht niedriger als die der Männer, was anzeigt, dass die Körpertemperatur der Frauen weniger variiert als die der Männer.

3. Beide Stichproben haben Werte für die standardisierte Schiefe und die standardisierte Wölbung im Bereich von -2 bis 2. Wie in Kapitel 10 erläutert, sind Werte in diesem Bereich konsistent mit der Hypothese, dass die Daten normalverteilt sind.

Ob die augenscheinliche Differenz zwischen den Temperaturen der Frauen und Männer statistisch signifikant ist, bleibt noch nachzuweisen.

11.3 Doppel-Histogramm

Das Häufigkeitsdiagramm liefert einen "Rücken-an-Rücken"-Vergleich der beiden Stichproben. Verwenden Sie die *Ergebnisfenster-Optionen* zur Neueinteilung der Klassenintervalle, so dass mit 25 Intervallen zwischen 96 und 101 Grad das folgende Diagramm generiert wird:

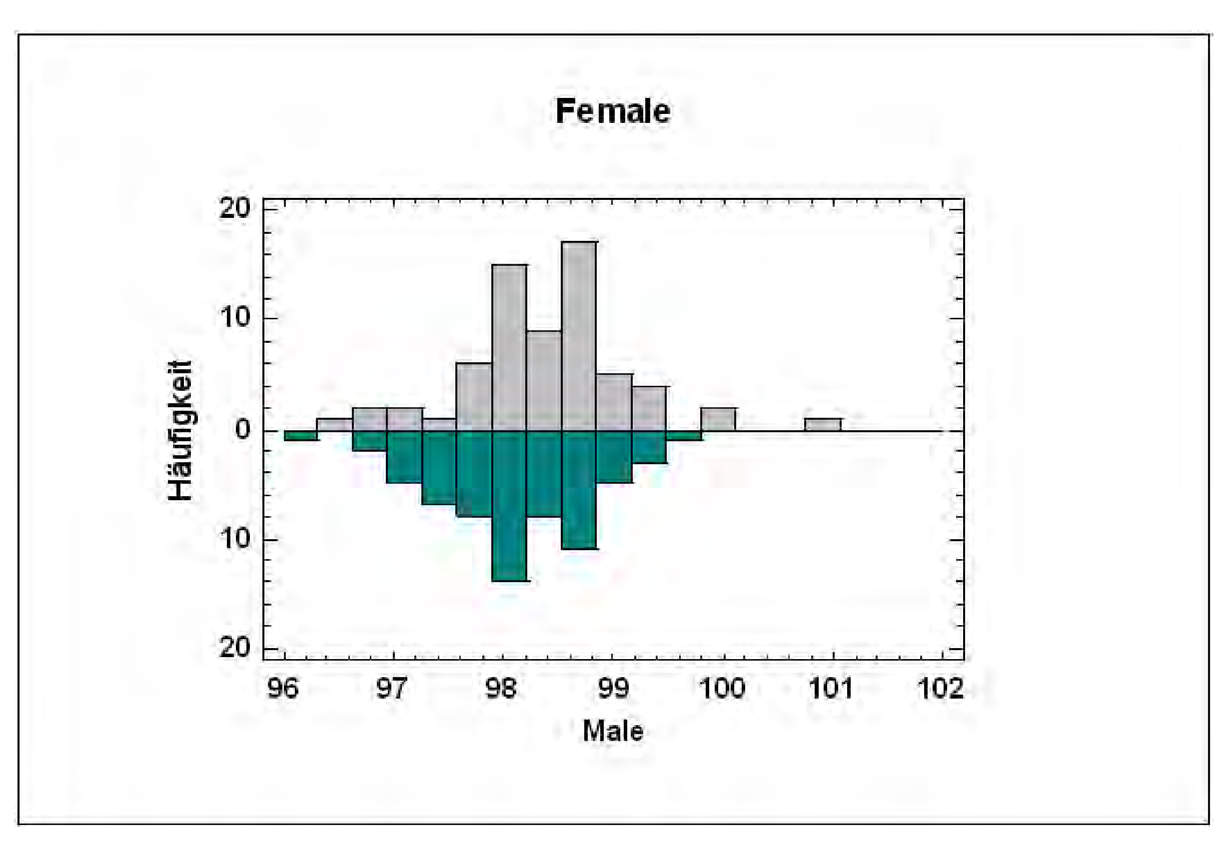


Abbildung 11.4: Doppel-Histogramm

Das Histogramm für die Frauen ist oberhalb der horizontalen Linie dargestellt. Das Histogramm für die Männer wird umgekehrt unterhalb der Linie angezeigt. Die Verteilungsformen sind ähnlich, mit einer möglichen Verschiebung der Verteilung für die Frauen nach rechts im Vergleich zu den Männern.

11.4 Zweifaches Box-Whisker-Plot

Im Analyse-Fenster werden für beide Stichproben Box-Whisker-Plots angezeigt. Wie in Kapitel 10 erläutert, umfasst die Box die mittlere Hälfte jeder Stichprobe. Die Ausläufer (Whisker) reichen bis zum größten und kleinsten Datenwert in jeder Stichprobe, mit Ausnahme derjenigen Punkte, die ungewöhnlich weit weg von der Box liegen. Die vertikale Linie in der Box stellt den Stichproben-Median dar, ein kleines Pluszeichen zeigt die Lage des Stichproben-Mittelwertes an.

In unserem Fall ist es besonders nützlich, zusätzlich die Median-Kerben darzustellen, die man bei den *Ergebnisfenster-Optionen* einstellen kann. Das Ergebnis wird nachfolgend angezeigt:

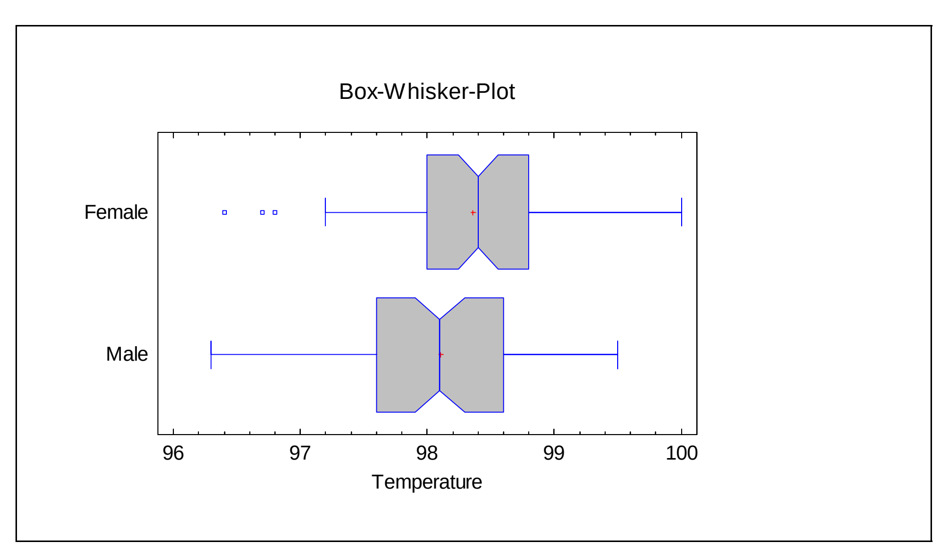


Abbildung 11..5: Zweifaches Box-Whisker-Plot mit Median-Kerben

Anhand dieses Diagramms wird offensichtlich:

1. Ein scheinbarer Abstand des Verteilungs-Zentrums für die Frauen rechts von dem der Männer. Beide Stichproben-Mittelwerte und -Mediane weisen ähnliche Differenzen auf.

2. Der Bereich, der durch die Werte der Frauen abgedeckt wird, ist breiter als der von den Männern, aber nur dann, wenn die kleinsten außerhalb liegenden Punkte mitbetrachtet werden.

3. Die Median-Kerbe für die Frauen überlappt geringfügig die der Männer. Die Kerben sind so eingestellt, dass von einem signifikanten Median-Unterschied bei der Standard-Irrtumswahrscheinlichkeit des Systems (in der Regel 5 %) ausgegangen werden kann, wenn *keine* Überlappung der beiden Kerben vorhanden ist. Ein formaler Test dafür wird in einem späteren Abschnitt beschrieben.

Anhand dieses Diagramms sieht es so aus, als wäre ein Mittelwert-Unterschied zwischen den beiden Stichproben vorhanden, die statistische Signifikanz dieser Differenz ist jedoch noch nicht belegt.

11.5 Vergleichen der Standardabweichungen

Der erste formale Test für den Vergleich zweier Stichproben ist der Test der Hypothese, dass die Standardabweichungen (σ) der Grundgesamtheiten, aus denen die Daten kommen, gleich sind, gegen die Alternativ-Hypothese, dass sie nicht gleich sind:

Null-Hypothese: $\sigma_1 = \sigma_2$

Alternativ-Hypothese: $\sigma_1 \neq \sigma_2$

Das ermöglicht uns, herauszufinden, ob die scheinbare Differenz in der Variabilität von Männern und Frauen statistisch signifikant ist oder ob sie im Rahmen der üblichen zufälligen Variabilität für Stichproben dieser Größe liegt.

Um diesen Test auszuführen, klicken Sie auf die Schaltfläche *Tabellen und Grafiken* auf der Analyse-Symbolleiste und wählen *Vergleich der Standardabweichungen* aus. Folgendes Ergebnis wird angezeigt:

Vergleich der Standardabweichungen für Temperature

	Gender=Female	*Gender=Male*
Standardabweichung	0,684262	0,698756
Varianz	0,468214	0,48826
FG	63	64

Varianzverhältnis = 0,958945

95,0% Konfidenzintervalle
Standardabweichung von Gender=Female: [0,582853; 0,828723]
Standardabweichung von Gender=Male: [0,595887; 0,844885]
Varianzverhältnis: [0,584028; 1,57609]

F-Test zum Vergleich von Standardabweichungen
Null-Hypothese: sigma1 = sigma2
Alternativ-Hypothese: sigma1 ungleich sigma2
F = 0,958945 p-Wert = 0,8684
Nullhypothese nicht ablehnen für alpha = 0,05.

Abbildung 11.6: Vergleich der Standardabweichungen für zwei Stichproben

Die wichtigsten Ergebnisse in dieser Tabelle sind rot hervorgehoben:

1. *Varianzverhältnis*: zeigt ein 95%-Konfidenzintervall für den Quotienten der Varianz für die Grundgesamtheit der Frauen σ_1^2 dividiert durch die Varianz der Grundgesamtheit der Männer σ_2^2. *Varianz* ist ein Maß für die Variabilität, berechnet als Quadrat der Standardabweichung. (HINWEIS: Vergleiche der Variabilität zwischen mehreren Stichproben werden üblicherweise auf Basis der Varianzen anstelle der Standardabweichungen durchgeführt, da sie die günstigeren mathematischen

Eigenschaften haben.) Das Intervall für σ_1^2 / σ_2^2 reicht von 0,58 bis 1,58. Das bedeutet, dass die Varianz für die Frauen irgendwo zwischen 58 % bis 158 % der Varianz für die Männer liegen kann. Diese mangelnde Genauigkeit ist typisch, wenn man versucht, die Variabilität anhand relativ kleiner Stichproben zu vergleichen.

2. Der *p-Wert* gehört zur *F-Statistik* über die oben genannte Hypothese. Ein *p*-Wert kleiner als 0,05 würde einen statistisch signifikanten Unterschied zwischen der Varianz für die Frauen und der für die Männer anzeigen, bei einer Irrtumswahrscheinlichkeit von 5 %. Da der *p*-Wert deutlich über 0,05 liegt, gibt es keinen Hinweis darauf, die Hypothese über gleiche Varianzen (und demzufolge gleiche Standardabweichungen) abzulehnen.

Folglich gibt es keinen statistisch signifikanten Unterschied zwischen der Variabilität für die Körpertemperaturen der Frauen und der für die Körpertemperaturen der Männer.

Es muss darauf hingewiesen werden, dass dieser Test sehr empfindlich reagiert, wenn die Voraussetzung verletzt ist, dass jede der Stichproben aus normalverteilten Grundgesamtheiten stammt. Die Voraussetzung scheint hier gegeben, wie mittels standardisierter Schiefe und standardisierter Wölbung gezeigt wurde.

11.6 Vergleichen der Mittelwerte

Der zweite Vergleich zwischen den beiden Stichproben testet die Hypothese, dass die Mittelwerte (μ) der beiden Grundgesamtheiten gleich sind:

Null-Hypothese: $\mu_1 = \mu_2$

Alternativ-Hypothese: $\mu_1 \neq \mu_2$

Um diesen Test auszuführen, klicken Sie wieder auf die Schaltfläche *Tabellen und Grafiken* und wählen *Vergleich der Mittelwerte*. Die Ergebnisse sind die folgenden:

```
Vergleich der Mittelwerte für Temperature
95,0% Konfidenzintervall für Mittelwert von: Gender=Female: 98,3563 +/- 0,170924   [98,1853; 98,5272]
95,0% Konfidenzintervall für Mittelwert von: Gender=Male: 98,1046 +/- 0,173144   [97,9315; 98,2778]
95,0% Konfidenzintervall für die Differenz zwischen den Mittelwerten
  unter der Annahme gleicher Varianz: 0,251635 +/- 0,240998   [0,0106371; 0,492632]

t-Test zum Vergleich der Mittelwerte
  Null-Hypothese: Mittelwert1 = Mittelwert2
  Alternativ-Hypothese: Mittelwert1 ungleich Mittelwert2
    unter der Annahme gleicher Varianz: t = 2,06616   p-Wert = 0,040846
  Nullhypothese ablehnen für alpha = 0,05.
```

Abbildung 11.7: Vergleich der Mittelwerte für zwei Stichproben

Die wichtigsten Ergebnisse in dieser Tabelle sind wiederum rot hervorgehoben:

1. *Differenz zwischen den Mittelwerten (unter der Annahme gleicher Varianzen)*: zeigt ein 95%-Konfidenzintervall an für den Mittelwert der Frauen-Population minus dem Mittelwert der Männer-Population. Das Intervall für μ_1 - μ_2 reicht von 0,01 bis 0,49, das bedeutet, dass der Mittelwert der weiblichen Körpertemperaturen zwischen 0,01 und 0,49 Grad über dem Mittelwert der männlichen Körpertemperaturen liegt.

2. Der *p-Wert* gehört zum t-Test über die oben genannte Hypothese. Da der *p*-Wert kleiner ist als 0,05, wird die Hypothese über die Gleichheit der Mittelwerte abgelehnt, und die beiden Populations-Mittelwerte werden als statistisch verschieden betrachtet bei einer Irrtumswahrscheinlichkeit von 5 %.

Beachten Sie, dass dieser Test unter der Voraussetzung gleicher Varianzen für beide Populationen durchgeführt wurde, was bereits mit der *F*-Statistik im vorhergehenden Abschnitt überprüft worden ist. Falls sich die Varianzen als signifikant verschieden herausgestellt hätten, könnte ein entsprechend angepasster t-Test angefordert werden, indem in den *Ergebnisfenster-Optionen* im Kontrollkästchen *Annahme gleicher Streuungen* das Häkchen entfernt wird.

Folglich sieht es so aus, dass die Frauen einer Grundgesamtheit mit einer im Mittel höheren Temperatur angehören als die der Männer.

11.7 Vergleichen der Mediane

Wenn der Verdacht besteht, dass die Daten Ausreißer enthalten, kann ein nichtparametrischer Test ausgeführt werden, der die Mediane anstelle der Mittelwerte vergleicht. Nichtparametrische Tests gehen *nicht* davon aus, dass die Daten normalverteilt sind, und sind robust gegenüber Ausreißern, sofern solche vorhanden sind.

Nach der Auswahl von *Vergleich der Mediane* aus der Dialogbox *Tabellen und Grafiken* wird der Mann-Whitney-(Wilcoxon)-*W*-Statistik durchgeführt. Bei diesem Test werden zunächst die beiden Stichproben vereinigt. Die zusammengeführten Daten werden dann in eine Rangfolge von 1 bis n_1+n_2 gebracht und die Originaldaten durch ihre entsprechende Rangzahl ersetzt. Die Testgröße *W* wird konstruiert, indem sie die mittleren Rangzahlen für die Beobachtungen beider Stichproben vergleicht:

Vergleich der Mediane für Temperature
Median von Stichprobe 1: 98,4
Median von Stichprobe 2: 98,1

Mann-Whitney (Wilcoxon) W-Test zum Vergleich der Mediane
  Null-Hypothese: median1 = median2
  Alternativ-Hypothese: median1 ungleich median2

  Mittlerer Rang von Stichprobe 1: 71,9219
  Mittlerer Rang von Stichprobe 2: 58,1846

  W = 1637,0   p-Wert = 0,0368312
  Nullhypothese ablehnen für alpha = 0,05.

Abbildung 11.8: Vergleich der Mediane für zwei Stichproben

Die Interpretation des Mann-Whitney-(Wilcoxon)-Tests erfolgt analog zu der des *t*-Tests, welcher im vorhergehenden Abschnitt beschrieben wurde. Ein kleiner *p*-Wert führt zu der Schlussfolgerung, dass die Mediane der beiden Grundgesamtheiten signifikant verschieden sind.

11.8 Quantil-Diagramm

Um den Unterschied zwischen den beiden Verteilungen zu veranschaulichen, können für die Stichproben nebeneinander liegende Quantil-Diagramme (Q-Plots) dargestellt werden, indem man *Quantil-Diagramm* in der Dialogbox *Grafiken* auswählt:

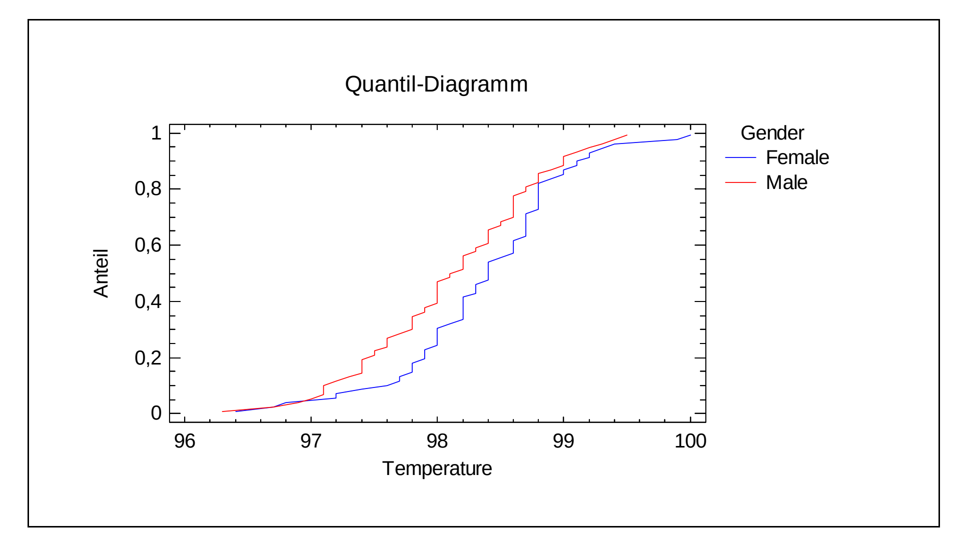


Abbildung 11.9: Nebeneinander liegende Quantil-Diagramme

Das Quantil-Diagramm veranschaulicht den Anteil an Daten in jeder Stichprobe, der unterhalb eines vorgegebenen Wertes von x liegt, als eine Funktion von x. Sofern die Stichproben aus derselben Grundgesamtheit stammen, sollten die Quantil-Diagramme eng beieinander liegen. Jede Verschiebung eines Q-Plots nach rechts oder links im Vergleich zu dem anderen weist auf einen Unterschied zwischen den beiden Medianen hin. Unterschiedliche Anstiege der Kurven zeigen eine Differenz zwischen den beiden Standardabweichungen an.

Im obigen Diagramm ist gut erkennbar, dass die Verteilung für die Frauen nach rechts verschoben ist, im Vergleich zu der Verteilung für die Männer. Die Anstiege sind jedoch ähnlich.

11.9 Kolmogorov-Smirnov-Test für zwei Stichproben

Ein weiterer Test, der verwendet werden kann, wenn die Voraussetzung der Normalverteilung nicht gegeben ist, ist der Kolmogorov-Smirnov-Test für zwei Stichproben. Dieser Test beruht auf der Berechnung des größten vertikalen Abstands zwischen den kumulativen Verteilungsfunktionen beider Stichproben, welcher näherungsweise gleich dem größten Abstand zwischen den beiden Q-Plots in Abbildung 11.9 ist. Falls der maximale Abstand groß genug ist, kann daraus geschlossen werden, dass die beiden Stichproben aus signifikant verschiedenen Grundgesamtheiten stammen.

Die Auswahl von *Kolmogorov-Smirnov-Test* aus der Dialogbox *Tabellen und Grafiken* führt zu folgendem Ergebnis:

Kolmogorov-Smirnov-Test für Temperature
Geschätzte Gesamt-Statistik DN = 0,242548
Zweiseitige K-S-Statistik für große Stichproben = 1,37737
Näherungsweiser p-Wert = 0,0449985

Abbildung 11.10: Ergebnis des Kolmogorov-Smirnov-Tests

Der maximale vertikale Abstand, der mit *DN* bezeichnet wird, ist für die Körpertemperatur-Daten näherungsweise gleich 0,24.

Der *p*-Wert wird benutzt, um herauszufinden, ob die beiden Verteilungen signifikant verschieden sind. Ein kleiner *p*-Wert führt zu der Schlussfolgerung, dass ein signifikanter Unterschied *existiert*. Da der *p*-Wert für die Stichproben-Daten kleiner als 0,05 ist, gibt es einen signifikanten Unterschied zwischen den Verteilungen für Männer und Frauen bei einer Irrtumswahrscheinlichkeit von 5 %.

Warnung: Wenn Daten stark gerundet sind, kann dieser Test ungeeignet sein, da die empirische Verteilungsfunktion (Cumulative Distribution Function, CDF) in großen Stufen springen kann. Wenn möglich, sollte man sich dann auf einen Vergleich von ausgewählten Verteilungsparametern wie Mittelwert, Median oder Standardabweichung stützen.

11.10 Quantil-Quantil-Diagramm

Ein letztes Diagramm, verfügbar bei Auswahl von *Quantil-Quantil-Diagramm* (Q-Q-Plot) in der Dialogbox *Grafiken*, stellt die geschätzten Quantile der einen Stichprobe gegen die Quantile der anderen Stichprobe dar:

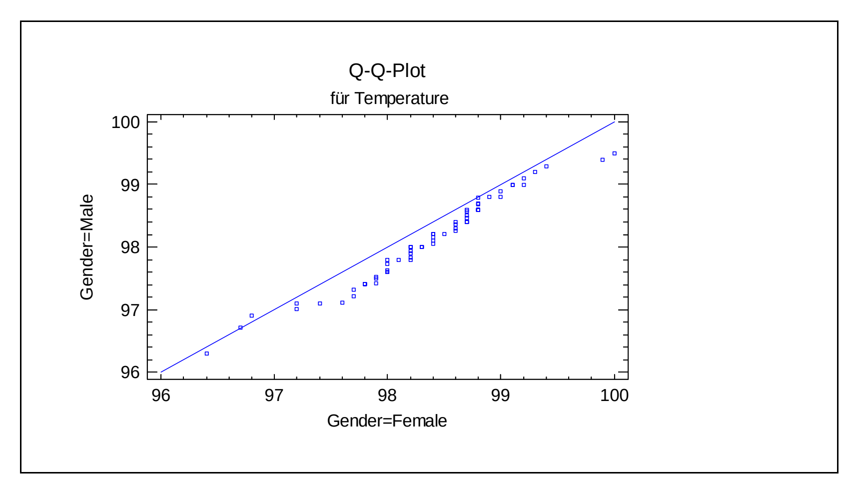


Abbildung 11.11: Q-Q-Plot der Körpertemperatur-Daten

In dieser Grafik gibt es einen Punkt zu jeder Beobachtung aus der kleineren der beiden Stichproben. Die x-Koordinate entspricht der Beobachtung und die y-Koordinate dem geschätzten Quantil der größeren Stichprobe. Sofern die Stichproben aus der gleichen Grundgesamtheit stammen, sollten die Punkte nahe an der Diagonalen liegen. Eine konstante Abweichung nach oben oder unten zeigt an, dass ein deutlicher Unterschied zwischen den Zentren der beiden Verteilungen besteht. Eine starke Abweichung des Anstieges der Punkte in den Rändern im Vergleich zum Anstieg der Diagonalen weist auf eine deutliche Differenz in der Variabilität hin. In unserem Fall ist der Unterschied zwischen den Verteilungen etwas komplexer als nur eine einfache Mittelwertverschiebung, da die Punkte für die hohen und niedrigen Temperaturen enger an der Linie liegen als für die mittleren Temperaturen. Es sieht so aus, als ob die Verteilung der weiblichen Temperaturen mehr im Zentrum konzentriert ist als die der männlichen.

Kapitel

12

Tutorial 3: Vergleichen mehrerer Stichproben

Vergleichen von Mittelwerten und Standardabweichungen, einfaktorielle ANOVA, ANOM und grafische Methoden.

Um Daten zu vergleichen, die in mehr als zwei Gruppen unterteilt sind, braucht man andere Techniken als die im vorhergehenden Kapitel verwendeten. Zum Beispiel wollen Sie vielleicht die Festigkeit von Gegenständen aus vier verschiedenen Materialien untersuchen. In einem typischen Experiment würden Sie 12 Teile aus jedem der vier Materialien herstellen, um diese zu vergleichen. Die folgenden Daten repräsentieren die Ergebnisse solch eines Experiments:

Material A	*Material B*	*Material C*	*Material D*
64,7	60,4	58,3	60,8
64,8	61,8	62,1	60,2
66,8	63,3	62,4	59,8
67,0	61,6	60,3	58,3
64,9	61,0	60,6	56,4
63,7	63,8	60,0	61,6
61,8	60,9	60,3	59,5
64,3	65,1	62,4	62,0
64,3	61,5	61,9	61,4
65,9	60,0	63,1	58,6
63,6	62,9	60,2	59,5
64,6	60,6	58,6	60,0

Es ist von besonderem Interesse herauszufinden, mit welchem der Materialien die stabilsten Teile produziert werden können sowie, welche Materialien sich statistisch voneinander unterscheiden.

Es gibt zwei Möglichkeiten, mehrere Stichproben in ein Datenblatt einzugeben:

1. Für jede Stichprobe eine separate Spalte verwenden.

2. Eine einzige Spalte für alle Daten verwenden und eine zweite Spalte erzeugen, welche die Kodes zur Identifizierung der jeweiligen Stichprobe, zu der die Daten gehören, enthält.

In unserem Beispiel verwenden wir die erste Variante. Die Daten für die Teile sind in den vier Spalten der Datei *widgets.sgd* enthalten. Diese Datei können Sie öffnen, indem Sie die Option *Öffnen – Datenquelle öffnen* im Menü *Datei* aufrufen.

12.1 Ausführen der Prozedur Vergleich mehrerer Stichproben

Die Prozedur *Vergleich mehrerer Stichproben* ist im Hauptmenü folgendermaßen verfügbar:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Vergleichen – Mehrere Stichproben – Vergleich mehrerer Stichproben.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Messbare Merkmale – Vergleich mehrerer Stichproben – Vergleich mehrerer Stichproben.*

In der ersten Dialogbox wird festgelegt, wie die Daten strukturiert sind:

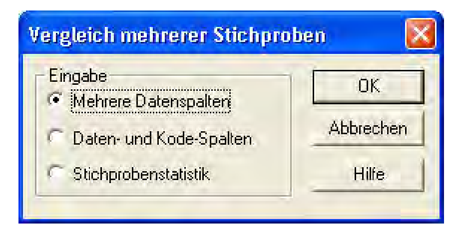


Abbildung 12.1: Erste Dialogbox für den Vergleich mehrerer Stichproben

In unserem Fall sind die Daten in mehreren Spalten des Datenblattes abgelegt.

Die zweite Dialogbox fragt nach den Namen der Spalten, welche die Daten enthalten:

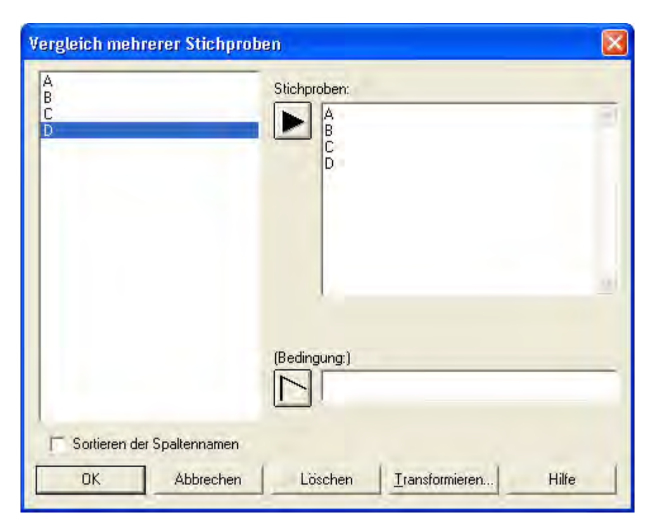
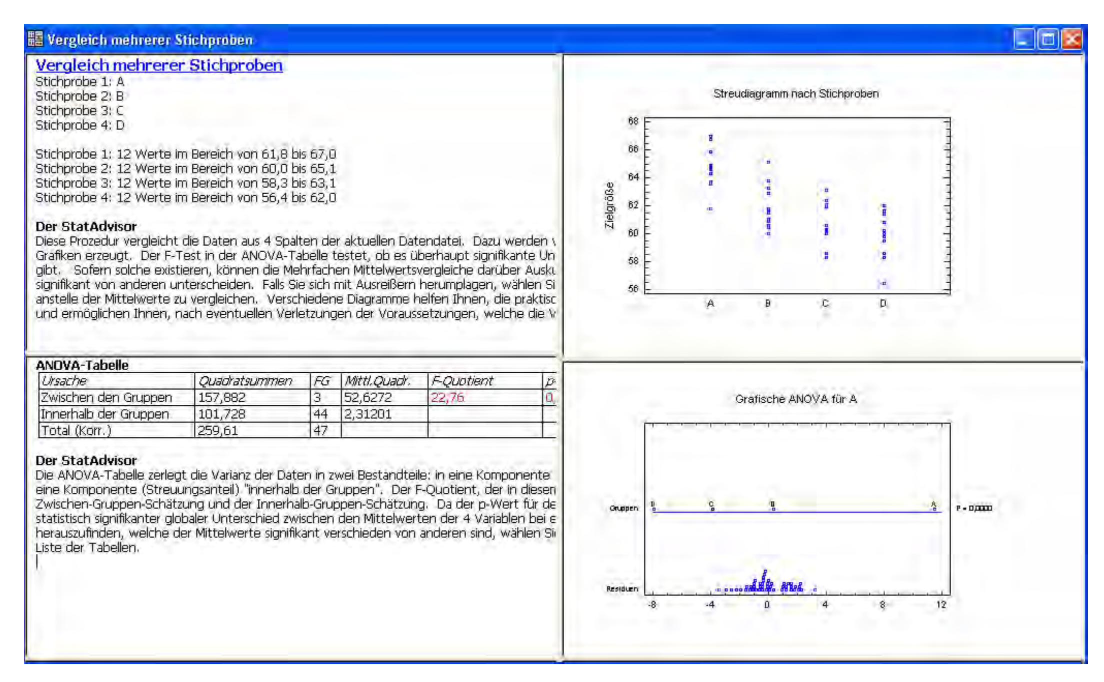


Abbildung 12.2: Dateneingabe-Dialogbox für den Vergleich mehrerer Stichproben

In der Beispieldatei sind die Beobachtungen in den vier Spalten mit den Namen A, B, C und D enthalten.

Nachdem Sie *OK* angeklickt haben, wird die Dialogbox *Tabellen und Grafiken* geöffnet. Für unser Beispiel können wir die Standardeinstellungen verwenden.

Wenn sich dann das Analyse-Fenster öffnet, sind zunächst vier Ergebnisfenster zu sehen:



Ursache	Quadratsummen	FG	Mittl.Quadr.	F-Quotient	p
Zwischen den Gruppen	157,882	3	52,6272	22,76	0,
Innerhalb der Gruppen	101,728	44	2,31201		
Total (Korr.)	259,61	47			

Abbildung 12.3: Analyse-Fenster für den Vergleich mehrerer Stichproben

Das obere linke Ergebnisfenster stellt eine Zusammenfassung der Analyse mit der Größe jeder Stichprobe und deren Wertebereichen dar. Das obere rechte Ergebnisfenster zeigt ein Streudiagramm der Daten, es folgt jetzt noch einmal vergrößert:

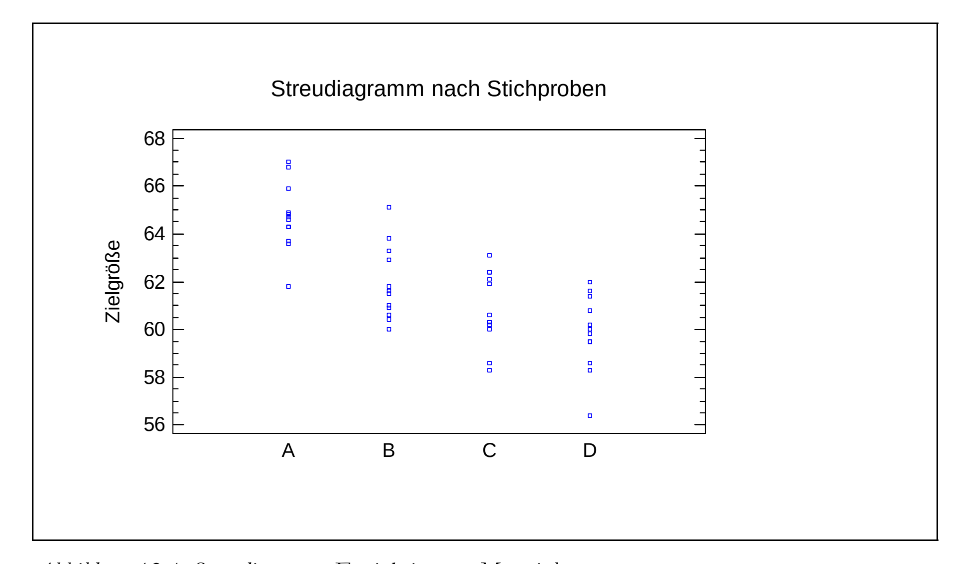


Abbildung 12.4: Streudiagramm Festigkeit gegen Material

Beachten Sie, dass einige der Beobachtungen übereinander dargestellt werden. Um dieses Problem zu beheben, klicken Sie doppelt in das Ergebnisfenster, um es zu vergrößern, und klicken dann auf die Schaltfläche *Auseinanderziehen* auf der Analyse-Symbolleiste. Durch Ziehen des oberen Gleiters ein wenig nach rechts fügen Sie eine kleine horizontale Verschiebung der Punkte ein:

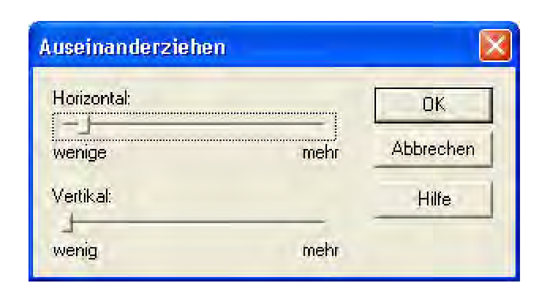


Abbildung 12.5: Dialogbox für das Auseinanderziehen

Diese zufällige Verschiebung um einen kleinen Wert in horizontaler Richtung macht die einzelnen Punkte leichter erkennbar:

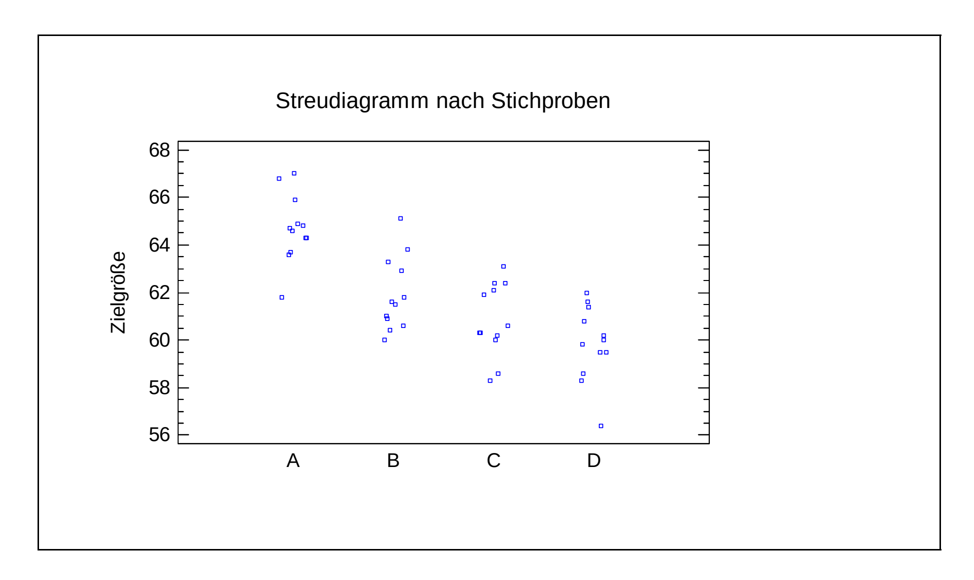


Abbildung 12.6: Streudiagramm nach dem Auseinanderziehen

Das Auseinanderziehen beeinflusst nur die Anzeige der Daten, nicht die Daten selbst oder irgendeine Berechnung aus den Daten.

12.2 Varianzanalyse

Der erste Schritt beim Vergleich mehrerer Stichproben ist üblicherweise eine einfaktorielle Varianzanalyse (ANOVA). Die ANOVA testet die Hypothese gleicher Mittelwerte der Grundgesamtheiten:

Null-Hypothese: $\mu_A = \mu_B = \mu_C = \mu_D$

Alternativ-Hypothese: die Mittelwerte sind nicht alle gleich

wobei μ_j den Mittelwert der Population repräsentiert, aus der die Stichprobe *j* gezogen wurde. Die Ablehnung der Null-Hypothese würde bedeuten, dass die Stichproben aus Grundgesamtheiten stammen, deren Mittelwerte nicht alle gleich sind.

Die Ergebnisse der ANOVA sind in der ANOVA-Tabelle enthalten, welche anfänglich im linken unteren Ergebnisfenster des Analyse-Fensters zu sehen wahr:

ANOVA-Tabelle

Ursache	*Quadratsummen*	*FG*	*Mittl.Quadr.*	*F-Quotient*	*p-Wert*
Zwischen den Gruppen	157,882	3	52,6272	22,76	0,0000
Innerhalb der Gruppen	101,728	44	2,31201		
Total (Korr.)	259,61	47			

Abbildung 12.7: ANOVA-Tabelle

Die Varianzanalyse zerlegt die Variabilität der Beobachtungswerte in zwei Bestandteile: in einen Anteil "Zwischen den Gruppen", der die Unterschiede zwischen den Teilen aus verschiedenen Materialien quantifiziert, und einen Anteil "Innerhalb der Gruppen", der die Unterschiede zwischen den Teilen aus dem gleichen Material quantifiziert. Falls die geschätzte Variabilität zwischen den Gruppen signifikant größer ist als die geschätzte Variabilität innerhalb der Gruppen, ist dies ein Hinweis darauf, dass die Gruppenmittelwerte nicht alle gleich sind.

Die Schlüsselgröße in *Abbildung 12.7* ist der *p*-Wert. Kleine *p*-Werte (kleiner als 0,05 bei einer Irrtumswahrscheinlichkeit von 5 %) führen zur Ablehnung der Hypothese über gleiche Mittelwerte. Im vorliegenden Beispiel gibt es kaum einen Zweifel, dass die Mittelwerte signifikant verschieden sind.

In der neuesten Ausgabe von Statistics for Experimenters von Box/Hunter/Hunter (John Wiley and Sons, 2005) präsentieren die Autoren ein neues Diagramm, das ANOVA-Ergebnisse in grafischer Form darstellt. Diese *Grafische ANOVA* wird standardmäßig im rechten unteren Ergebnisfenster angezeigt:

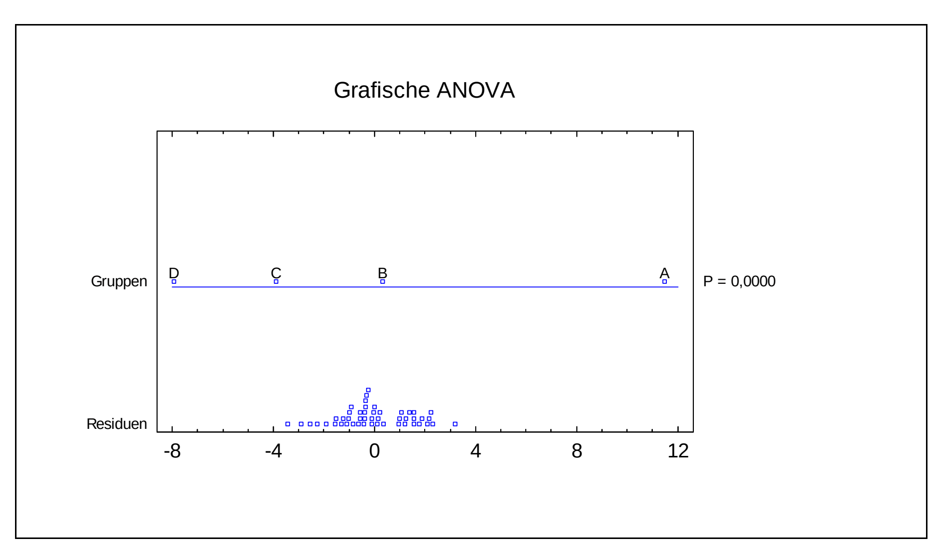


Abbildung 12.8: Grafische ANOVA

Entlang der Grundlinie der Grafik findet man ein Dot-Diagramm der *Residuen* des Modells. In der einfaktoriellen ANOVA sind die Residuen gleich der Differenz zwischen jeder Beobachtung und dem zugehörigen Gruppenmittelwert. In unserem Beispiel ist die beobachtete Variabilität in den Residuen ein Zeiger für die natürliche Variabilität unter den Teilen, die aus dem gleichen Material gefertigt wurden. Oberhalb der Mittellinie werden skalierte Abweichungen der Gruppenmittelwerte vom Gesamtmittelwert aller n = 48 Beobachtungen aufgetragen. Diese Gruppenabweichungen sind so skaliert, dass ihre Variabilität mit der von den Residuen verglichen werden kann. Alle Gruppen, deren Punkte zu weit abseits sind, um zu einer Verteilung mit einer ähnlichen Streubreite wie die der Residuen zu gehören, korrespondieren wahrscheinlich mit anderen Populationen.

In Abbildung 12.8 scheint die Gruppe A klar von den anderen Gruppen getrennt zu sein. Die Trennung der anderen drei Mittelwerte ist weniger deutlich. Ein formaler Vergleich der vier Stichprobenmittelwerte wird im nächsten Abschnitt beschrieben.

12.3 Vergleichen der Mittelwerte

Falls der p-Wert in der ANOVA-Tabelle klein ist, sollten die Stichprobenmittelwerte daraufhin untersucht werden, welche der Mittelwerte signifikant voneinander verschieden sind. Für diesen Zweck ist das *Mittelwert-Diagramm* gut geeignet, das man in der Dialogbox *Tabellen und Grafiken* aufrufen kann:

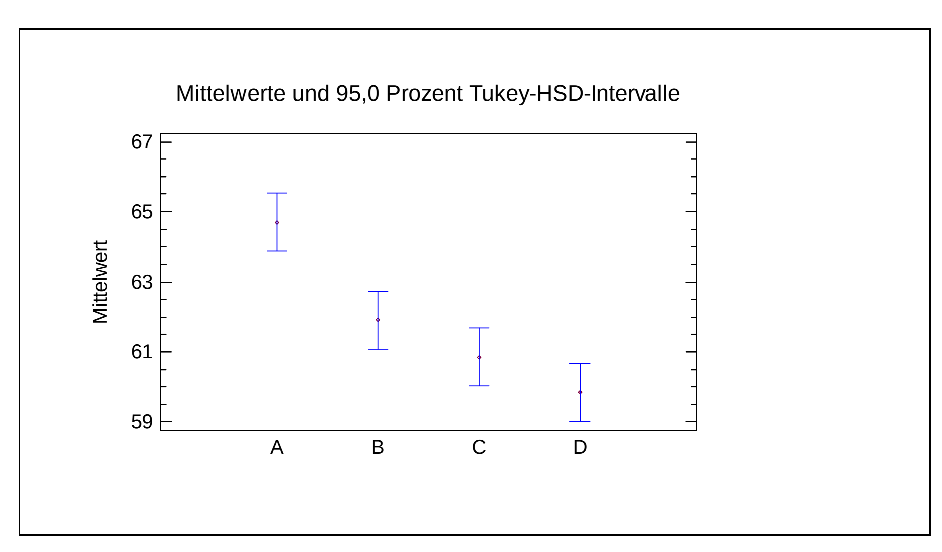


Abbildung 12.9: Mittelwert-Diagramm

Das Mittelwert-Diagramm zeigt jeden Stichprobenmittelwert zusammen mit einem Schätzintervall an. Die Interpretation dieses Intervalls hängt vom ausgewählten Intervall-Typ ab, den man mittels *Ergebnisfenster-Optionen* einstellen kann. Die am häufigsten verwendeten Intervalle sind:

1. *Fishers LSD-(Least Significant Difference)-Intervalle*: Diese Intervalle sind so aufgebaut, dass man ein einzelnes Stichproben-Paar auswählen und deren Mittelwerte als signifikant verschieden betrachten kann, wenn sich die Intervalle in vertikaler Richtung nicht überlappen. Wenn die Wahrscheinlichkeit, zwei Mittelwerte fälschlicherweise als verschieden zu deklarieren, auf 5 % festgelegt ist, kann der Vergleich mehrerer Mittelwert-Paare auf diesem Niveau zu einem Fehler für das Gesamtverfahren führen, der wesentlich größer ist als diese 5 %.

2. *Tukeys HSD-(Honestly Significant Difference)-Intervalle*: Diese Intervalle sind so konzipiert, dass sie den Fehler von 5 % für das Gesamtverfahren einhalten. Bei Verwendung der Tukey-Methode werden Sie also höchstens in 5 % der Fälle Mittelwertpaare fälschlicherweise als signifikant verschieden deklarieren.

Die Intervalle in *Abbildung 12.9* sind Tukey-Intervalle. Da das Intervall für Stichprobe A mit keinem der anderen Intervalle überlappt, ist der Mittelwert von A signifikant verschieden von den anderen drei Mittelwerten. Ebenso ist B signifikant verschieden von D, da ihre Intervalle nicht überlappen. C jedoch ist weder zu B noch zu D signifikant verschieden.

Dieselbe Analyse kann in Tabellenform angezeigt werden, indem man *Mehrfache Mittelwertsvergleiche* in der Dialogbox *Tabellen und Grafiken* aufruft:

Mehrfache Mittelwertvergleiche

Methode: 95,0 Prozent Tukey HSD

	Anzahl	*Mittelwert*	*Homogene Gruppen*
D	12	59,8417	X
C	12	60,85	XX
B	12	61,9083	X
A	12	64,7	X

Kontrast	*Sig.*	*Differenz*	*+/- Grenzen*
A - B	*	2,79167	1,65755
A - C	*	3,85	1,65755
A - D	*	4,85833	1,65755
B - C		1,05833	1,65755
B - D	*	2,06667	1,65755
C - D		1,00833	1,65755

* bezeichnet einen statistisch signifikanten Unterschied.

Abbildung 12.10: Mehrfache Mittelwertsvergleiche

Der untere Bereich der Ausgabe zeigt alle Mittelwert-Paare an. In der Spalte *Differenz* wird die Differenz aus dem Stichprobenmittelwert der ersten Gruppe und dem der zweiten Gruppe berechnet. Die Spalte +/- *Grenzen* beinhaltet ein Schätzintervall für diese Differenz. Jedes Paar, für das der Absolutwert der Differenz die Grenzen überschreitet, ist statistisch signifikant mit der gewählten Irrtumswahrscheinlichkeit und wird durch ein * in der Spalte Signifikanz (*Sig.*) gekennzeichnet. In unserem Beispiel weisen vier der sechs Mittelwert-Paare signifikante Differenzen auf.

Der obere Bereich in der Anzeige ordnet die Stichproben in homogene Gruppen ein, die als Spalten mit Kreuzen (X) angezeigt werden. Eine homogene Gruppe ist eine Gruppe, in der keine signifikanten Differenzen auftreten. In unserem Fall bildet Stichprobe A eine Gruppe für sich, da sie signifikant verschieden von allen anderen ist. Stichprobe C fällt in zwei Gruppen, in eine mit B und in eine mit D. Es wären mehr Daten erforderlich, um zu unterscheiden, zu welcher Gruppe die Stichprobe C tatsächlich gehört.

12.4 Vergleichen der Mediane

Wenn der Verdacht besteht, dass die Daten Ausreißer enthalten, kann ein nichtparametrisches Verfahren alternativ zu einer Standard-Varianzanalyse ausgeführt werden, indem man *Kruskal-Wallis-* und *Friedman-Tests* in der Dialogbox *Tabellen und Grafiken* auswählt. Diese Tests vergleichen die Stichproben-Mediane anstelle der Mittelwerte:

Null-Hypothese: alle Mediane sind gleich

Alternativ-Hypothese: die Mediane sind nicht alle gleich

Die Testmethode kann man über die *Ergebnisfenster-Optionen* auswählen. Es werden zwei Tests angeboten:

1. *Kruskal-Wallis-Test* – geeignet, wenn jede Spalte eine zufällige Stichprobe aus der jeweiligen Grundgesamtheit enthält. In solch einem Fall haben die Zeilen eigentlich keine Bedeutung.

2. *Friedman-Test* – geeignet, wenn jede Zeile einen Block repräsentiert, d. h. die Stufe irgendeiner anderen Variablen. Typische Blockvariablen sind die Wochentage, Herstellerstandorte oder Schichten.

Im Beispiel haben die Zeilen keine Bedeutung, deshalb ist der Kruskal-Wallis-Test angemessen:

Kruskal-Wallis-Test

	St.pr.umfang	*Mittlerer Rang*
A	12	40,7917
B	12	25,7917
C	12	19,25
D	12	12,1667

Teststatistik = 27,3735 p-Wert = 0,00000491592

Abbildung 12.11: Kruskal-Wallis-Test

Der wichtigste Eintrag in der obigen Tabelle ist der *p*-Wert. Da der *p*-Wert sehr klein ist (kleiner als 0,05), wird die Hypothese der Mediangleichheit abgelehnt.

Man kann auch die Median-Paare vergleichen, indem man *Box-Whisker-Plots* in der Dialogbox *Tabellen und Grafiken* auswählt und mittels der *Ergebnisfenster-Optionen* Mediankerben hinzufügt:

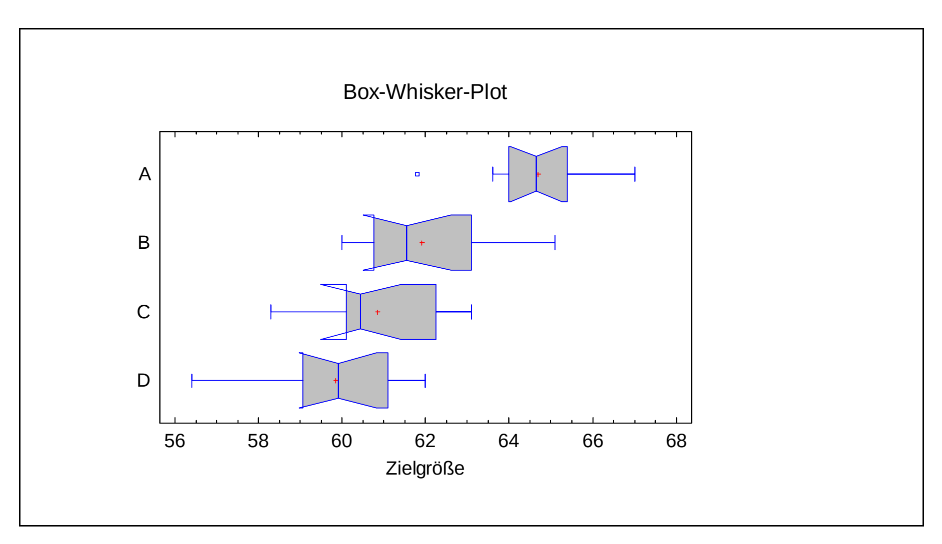


Abbildung 12.12: Box-Whisker-Plots mit Mediankerben

Der Bereich, der durch jede der Kerben überdeckt wird, verkörpert die Unsicherheit, die mit der Schätzung des Stichproben-Medians verbunden ist. Die Kerben sind so eingestellt, dass von einem signifikanten Median-Unterschied zwischen jeweils zwei Stichproben ausgegangen werden kann, bei der Standard-Irrtumswahrscheinlichkeit des Systems (in der Regel 5 %), wenn *keine* Überlappung der beiden Kerben vorhanden ist. Im obigen Diagramm überlappen die Kerben für die Stichproben B, C und D, nur der Median für Stichprobe A ist signifikant höher als die anderen drei.

> ANMERKUNG: Das „gefaltete“ Aussehen von zwei der Boxen in Abbildung 12.12 resultiert daraus, dass die Kerben über die Box hinausgehen.

12.5 Vergleichen der Standardabweichungen

Man kann auch die Hypothese über gleiche Standardabweichungen testen:

Null-Hypothese: $\sigma_A = \sigma_B = \sigma_C = \sigma_D$

Alternativ-Hypothese: die Standardabweichungen sind nicht alle gleich

indem man *Varianzprüfung* in der Dialogbox *Tabellen und Grafiken* auswählt:

Varianzprüfung		
	Test	*p-Wert*
Levenes	0,143286	0,933432

Abbildung 12.13: Vergleich der Stichproben-Varianzen

Einer von vier Tests wird angezeigt, in Abhängigkeit von der Auswahl in *Ergebnisfenster-Optionen*. Drei der verfügbaren Tests, einschließlich des Levenes-Tests, geben einen *p*-Wert aus. Ein *p*-Wert kleiner als 0,05 führt zur Ablehnung der Null-Hypothese bei einer Irrtumswahrscheinlichkeit von 5 %. In unserem Fall sind die Standardabweichungen nicht signifikant voneinander verschieden, da der *p*-Wert deutlich über 0,05 liegt.

Zusammengefasst sieht es so aus, dass die mittlere Festigkeit für unterschiedliche Materialien verschieden ist. Die Variabilität zwischen den Teilen aus dem gleichen Material ist jedoch für alle vier Materialien in etwa gleich.

12.6 Residuen-Diagramme

Immer, wenn ein statistisches Modell an Daten angepasst wird, ist die Untersuchung der Residuen des angepassten Modells von Bedeutung. In dieser Analyse wird ein Residuum für jeden der $n = 48$ Gegenstände definiert als die Differenz zwischen der Festigkeit des Gegenstandes und der mittleren Festigkeit für alle Teile, die aus dem gleichen Material gefertigt wurden.

Die Dialogbox *Grafiken* enthält eine Option für automatisch generierte Residuen-Diagramme. In Abhängigkeit von der Auswahl in *Ergebnisfenster-Optionen* können Residuen nach Faktorstufen (Gruppen), gegen die Modellwerte oder gegen die Reihenfolge (Zeilennummer in der Datendatei) gezeichnet werden. Das folgende Diagramm zeigt die Residuen gegen die Modellwerte von *strength*:

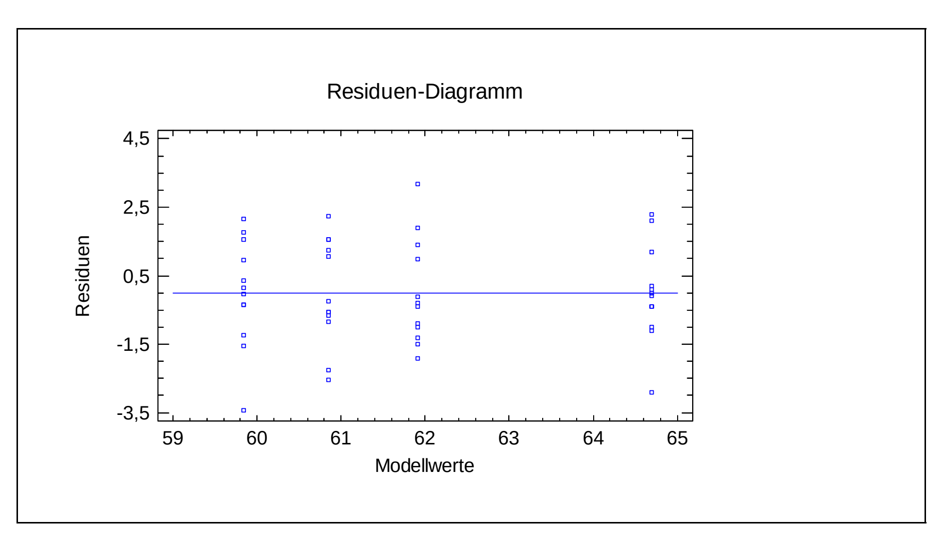


Abbildung 12.14: Diagramm der Residuen gegen die Modellwerte von Strength

Mit Hilfe dieser drei Diagrammtypen sollten Sie nach folgenden Dingen schauen:

1. *Ausreißer* – isolierte Residuen, weit weg von allen anderen. Derartige Punkte würden eine weitere Untersuchung erfordern, um herauszufinden, ob es eine zuordenbare Ursache zur Erklärung dieses ungewöhnlichen Verhaltens gibt.

2. *Heteroskedastizität* – systematische Veränderung der Varianz mit dem Anwachsen bzw. Abfallen der Zielgrößenwerte. Das resultiert typischerweise in einem trichterförmigen Erscheinungsbild der Residuen und kann eine Transformation der Originalbeobachtungswerte erforderlich machen, indem vor Ausführen der Analyse der Logarithmus gebildet wird. Verfahren wie *Mehrfacher Mittelwertvergleich* arbeiten nicht korrekt, wenn die Innere-Gruppen-Variabilität zwischen den Gruppen stark unterschiedlich (nicht homogen) ist.

Sofern gewünscht, können die Residuen als neue Spalte in ein Datenblatt gespeichert werden, indem die Schaltfläche *Ergebnisse speichern* auf der Analyse-Symbolleiste angeklickt wird.

12.7 Mittelwertanalyse-Diagramm (ANOM)

Ein etwas anderer Weg zum Vergleich verschiedener Mittelwerte ist die Verwendung des *ANOM-Diagramms (Mittelwertanalyse-Diagramms)*, ebenfalls verfügbar in der Dialogbox *Tabellen und Grafiken*:

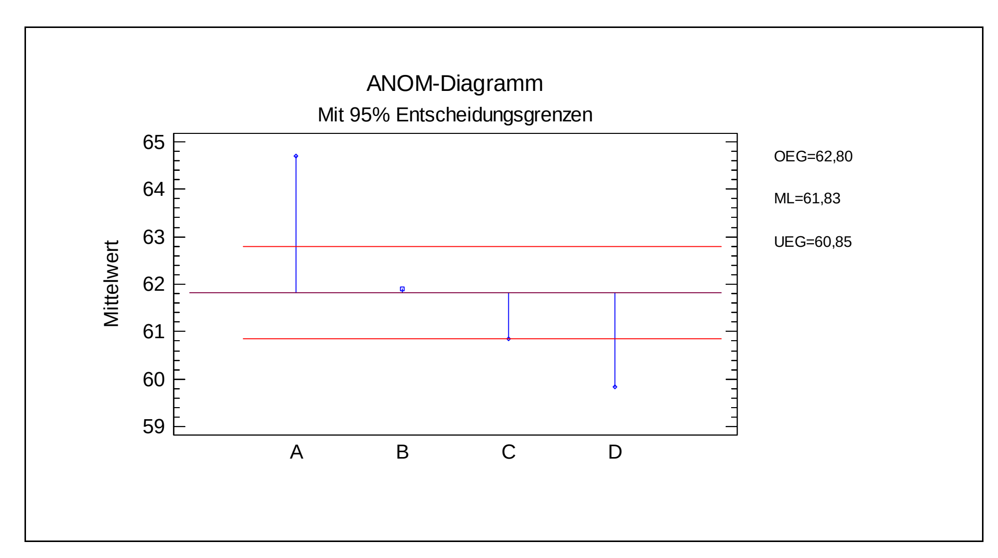


Abbildung 12.15: ANOM-Diagramm

Vom Aussehen her ähnlich wie eine Regelkarte, zeigt dieses Diagramm alle Stichprobenmittelwerte mit einer vertikalen Linie, die zum Gesamtmittelwert aller Beobachtungen reicht. Ober- und unterhalb des Gesamtmittelwertes sind Entscheidungsgrenzen eingezeichnet. Jeder Stichprobenmittelwert, der außerhalb der Entscheidungsgrenzen liegt, wird als signifikant verschieden vom Gesamtmittelwert betrachtet.

Die Interpretation in unserem Fall ist folgende: Die Gegenstände aus der Stichprobe A besitzen eine signifikant höhere Festigkeit als der Durchschnitt, während die Teile der Stichproben C und D signifikant weniger fest sind als der Durchschnitt. Diese Art der Interpretation kann mitunter sehr hilfreich sein.

Kapitel

13

Tutorial 4: Regressionsanalyse

Anpassen linearer und nichtlinearer Modelle, Auswählen des besten Modells, Darstellen der Residuen und Anzeigen der Ergebnisse.

Die Gruppe der Prozeduren zur Anpassung statistischer Regressionsmodelle gehört zu den am meisten verwendeten Teilen von STATGRAPHICS Centurion XVI. In einem Regressionsmodell wird eine Zielgröße Y als eine Funktion von einer oder mehreren erklärenden Variablen (Einflussgrößen) X sowie einer Störung aufgefasst. In vielen Fällen (aber nicht immer) ist dieser funktionale Form linear in den unbekannten Koeffizienten, so dass das Modell folgendermaßen beschrieben werden kann:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 x_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_k X_{k,i} + \varepsilon_i$$

wobei der Index *i* die *i*. Beobachtung in der Stichprobe, die β's unbekannte Modellkoeffizienten und ε eine zufällige Schwankung sind. Letztere wird in der Regel als normalverteilt mit dem Mittelwert 0 und der Standardabweichung σ vorausgesetzt.

In einer gegebenen Datenmenge mit einer Zielgröße Y und einer oder mehreren Einflussgrößen besteht das Ziel der Regressionsanalyse darin, ein Modell zu finden, das:

1. die zwischen den Variablen existierende Beziehung beschreibt und es dann gestattet, Y für bekannte Werte von X vorherzusagen.

2. nicht mehr X-Variablen enthält als notwendig, um gute Vorhersagen zu generieren.

Die zweite Eigenschaft wird manchmal als *Parsimonität* (Sparsamkeit) bezeichnet. In der Regel schneiden Modelle mit einer kleinen Menge gut ausgewählter Einflussgrößen in der Praxis am besten ab.

Dieses Kapitel betrachtet verschiedene Arten von Regressionsmodellen. Im Beispiel soll die Größe *MPG City* (Anzahl Meilen pro Gallone Benzin bei Stadtfahrten) für die Autos in der Datei *93cars.sgd* als Zielgröße Y dienen. Das Ziel besteht darin, mit Hilfe der anderen Variablen dieser Datei ein Modell zu bilden, welches die Meilen pro Gallone Benzin für ein Auto bestens vorhersagen kann.

13.1 Korrelationsanalyse

Ein sinnvoller Beginn für die Bildung eines Regressionsmodells ist die Prozedur *Analyse mehrerer Variablen*. Diese Analyse findet man im Hauptmenü folgendermaßen:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beschreiben – Numerische Daten – Analyse mehrerer Variablen.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Messbare Daten – Multivariate Methoden – Analyse mehrerer Variablen.*

Diese Analyse zeigt zu Beginn folgende Dateneingabe-Dialogbox an:

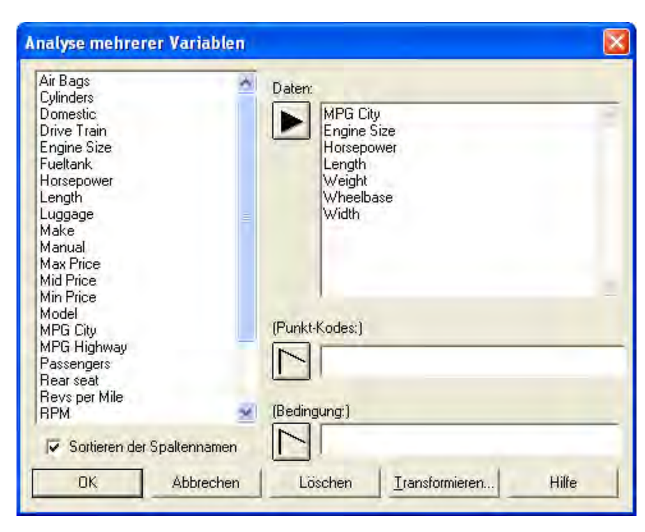
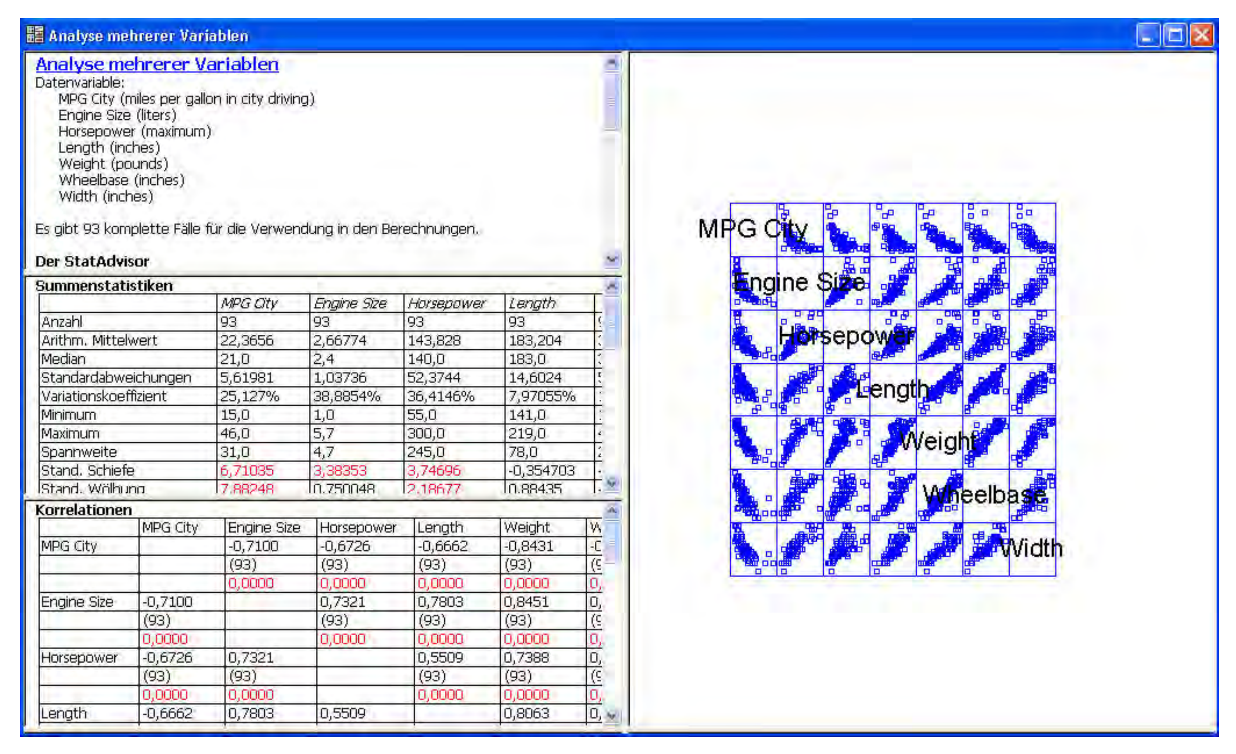


Abbildung 13.1: Dateneingabe-Dialogbox für Analyse mehrerer Variablen

Sechs mögliche erklärende Variable wurden ausgewählt, zusätzlich zu *MPG City*. Die potentiellen Einflussgrößen sind:

- X_1: *Engine Size* (Liters)
- X_2: *Horsepower* (Maximum)
- X_3: *Length* (Inches)
- X_4: *Weight* (Pounds)
- X_5: *Wheelbase* (Inches)
- X_6: *Width* (Inches)

Nach dem Anklicken von *OK* erscheinen die *Optionen*-Dialogbox, die *Tabellen und Grafiken*-Dialogbox und anschließend das neue Analyse-Fenster:



	MPG City	Engine Size	Horsepower	Length
Anzahl	93	93	93	93
Arithm. Mittelwert	22,3656	2,66774	143,828	183,204
Median	21,0	2,4	140,0	183,0
Standardabweichungen	5,61981	1,03736	52,3744	14,6024
Variationskoeffizient	25,127%	38,8854%	36,4146%	7,97055%
Minimum	15,0	1,0	55,0	141,0
Maximum	46,0	5,7	300,0	219,0
Spannweite	31,0	4,7	245,0	78,0
Stand. Schiefe	6,71035	3,38353	3,74696	-0,354703
Stand. Wölbung	7,88248	0,750048	2,18677	0,88435

	MPG City	Engine Size	Horsepower	Length	Weight	W
MPG City		-0,7100	-0,6726	-0,6662	-0,8431	-C
		(93)	(93)	(93)	(93)	(9
		0,0000	0,0000	0,0000	0,0000	0,
Engine Size	-0,7100		0,7321	0,7803	0,8451	0,
	(93)		(93)	(93)	(93)	(9
	0,0000		0,0000	0,0000	0,0000	0,
Horsepower	-0,6726	0,7321		0,5509	0,7388	0,
	(93)	(93)		(93)	(93)	(9
	0,0000	0,0000		0,0000	0,0000	0,
Length	-0,6662	0,7803	0,5509		0,8063	0,

Abbildung 13.2: Analyse-Fenster der Analyse für mehrere Variablen

Das obere linke Ergebnisfenster zeigt die eingegebenen Variablen an und das mittlere linke Ergebnisfenster die Summenstatistiken. In der Datendatei gibt es 93 Zeilen mit kompletten Informationen in allen zu analysierenden Variablen (keine fehlenden Werte vorhanden).

Das Matrix-Diagramm im rechten Ergebnisfenster stellt x-y-Diagramme für alle Variablenpaare dar.

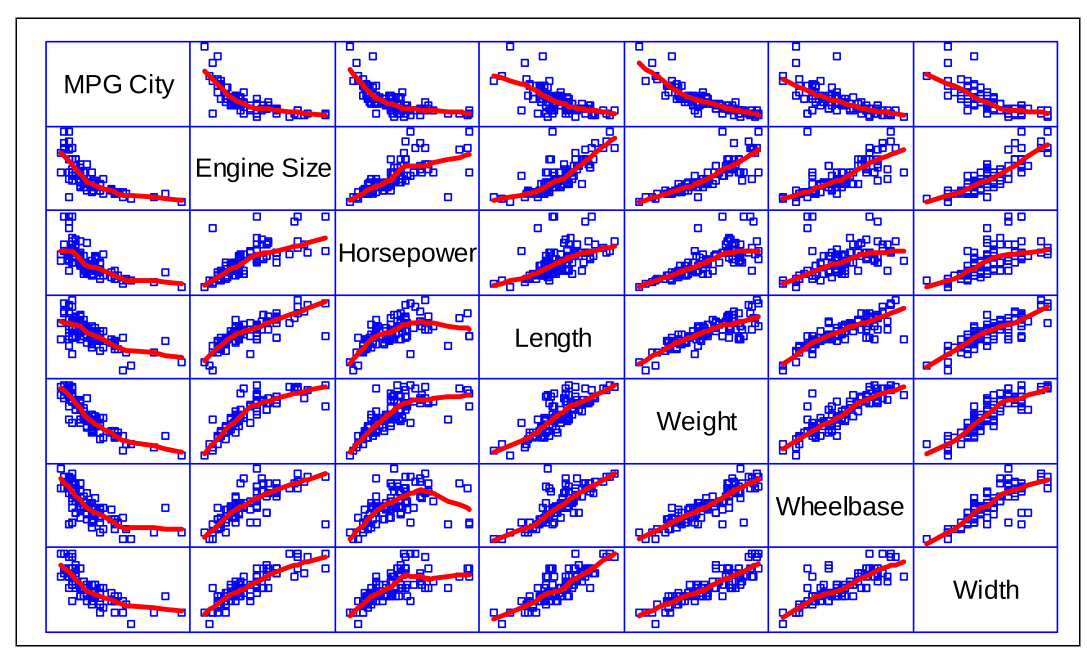


Abbildung 13.3: Matrix-Diagramm mit zusätzlicher Glättung

Zur Interpretation dieses Diagramms suchen wir nach einer Variablen-Beschriftung wie z.B. *MPG City*. Diese Variable wird in jedem kleinen Diagramm in der Zeile *MPG City* an der vertikalen Achse aufgetragen und in jedem kleinen Diagramm in der Spalte *MPG City* an der horizontalen Achse. Jedes Variablenpaar wird somit zweimal angezeigt, einmal oberhalb und einmal unterhalb der Diagonalen.

Zusätzlich kann man robuste LOWESS-Glätter in die obige Abbildung einfügen, indem man das Ergebnisfenster maximiert und die Schaltfläche *Glätten/Rotieren* auf der Analyse-Symbolleiste anklickt. Von besonderem Interesse ist die oberste Zeile des Matrix-Diagramms, die *MPG City* gegen alle sechs potentiellen erklärenden Variablen darstellt. Jede der Variablen steht in deutlicher Beziehung zu *MPG City*, manche in einer nichtlinearen Form. Außerdem ist erhebliche Multikollinearität (Korrelation zwischen den erklärenden Variablen) vorhanden, was darauf hindeutet, dass viele verschiedene Variablenkombinationen gleichermaßen gut geeignet sein können, um Y vorherzusagen.

Die Tabelle links unten zeigt die Matrix der geschätzten Korrelationskoeffizienten für jedes Variablenpaar an.

Korrelationen

	MPG City	Engine Size	Horsepower	Length	Weight	Wheelbase	Width
MPG City		-0,7100	-0,6726	-0,6662	-0,8431	-0,6671	-0,7205
		(93)	(93)	(93)	(93)	(93)	(93)
		0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Engine Size	-0,7100		0,7321	0,7803	0,8451	0,7325	0,8671
	(93)		(93)	(93)	(93)	(93)	(93)
	0,0000		0,0000	0,0000	0,0000	0,0000	0,0000
Horsepower	-0,6726	0,7321		0,5509	0,7388	0,4869	0,6444
	(93)	(93)		(93)	(93)	(93)	(93)
	0,0000	0,0000		0,0000	0,0000	0,0000	0,0000
Length	-0,6662	0,7803	0,5509		0,8063	0,8237	0,8221
	(93)	(93)	(93)		(93)	(93)	(93)
	0,0000	0,0000	0,0000		0,0000	0,0000	0,0000
Weight	-0,8431	0,8451	0,7388	0,8063		0,8719	0,8750
	(93)	(93)	(93)	(93)		(93)	(93)
	0,0000	0,0000	0,0000	0,0000		0,0000	0,0000
Wheelbase	-0,6671	0,7325	0,4869	0,8237	0,8719		0,8072
	(93)	(93)	(93)	(93)	(93)		(93)
	0,0000	0,0000	0,0000	0,0000	0,0000		0,0000
Width	-0,7205	0,8671	0,6444	0,8221	0,8750	0,8072	
	(93)	(93)	(93)	(93)	(93)	(93)	
	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	

Korrelation
(Stichprobengröße)
p-Wert

Abbildung 13.4: Korrelationsmatrix

Die Tabelle zeigt die Korrelationskoeffizienten für jedes Variablenpaar, die Anzahl der für die Schätzung vorwendeten Beobachtungen und einen *p*-Wert. Der Korrelationskoeffizient *r* ist eine Zahl zwischen –1 und +1, welche die Stärke der linearen Abhängigkeit zwischen zwei Variablen angibt. Je näher der Korrelationskoeffizient an –1 oder +1 liegt, umso stärker ist der Zusammenhang. Das Vorzeichen weist auf die Richtung des Zusammenhangs hin. Ein positiver Wert bedeutet, dass Y zusammen mit X ansteigt, ein negativer Wert bedeutet ein gegenläufiges Verhalten von Y und X.

Um herauszufinden, ob zwei Variablen signifikant linear abhängig sind, wird zu jedem Korrelationskoeffizienten ein *p*-Wert berechnet. Für jedes Variablenpaar mit einem *p*-Wert kleiner als 0,05 wird die Hypothese über die lineare Unabhängigkeit abgelehnt, bei einer Irrtumswahrscheinlichkeit von 5 %, d.h. die Variablen sind linear abhängig.

Die oberste Zeile der Korrelationsmatrix enthält die Korrelationskoeffizienten zwischen *MPG City* und den 6 Einflussgrößen. Die stärkste Korrelation besteht mit -0,8431 zu *Weight*. Das

negative Vorzeichen impliziert, dass *MPG City* mit wachsendem *Weight* abnimmt, was nicht allzu sehr überrascht.

13.2 Einfache Regression

Das erste statistische Modell, was angepasst werden soll, ist eine Gerade der Form:

$$MPG\ City = \beta_0 + \beta_1 Weight + \varepsilon$$

In dieser Gleichung stellt β_1 den Anstieg der Geraden dar (in Einheiten von Meilen je Gallone Benzin pro Pfund), während β_0 das Y-Absolutglied ist. Zur Anpassung dieses Modells:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beziehungen – Ein Faktor – Einfache Regression.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Verbessern – Regressionsanalyse – Ein Faktor – Einfache Regression.*

Die Dateneingabe-Dialogbox sollte folgendermaßen aussehen:



Abbildung 13.5: Dateneingabe-Dialogbox der Einfachen Regression

Nach den *Optionen* und der Dialogbox *Tabellen und Grafiken* wird das neue Analyse-Fenster erstellt. Es besteht aus vier Ergebnisfenstern mit Informationen zum angepassten Modell und den Residuen:

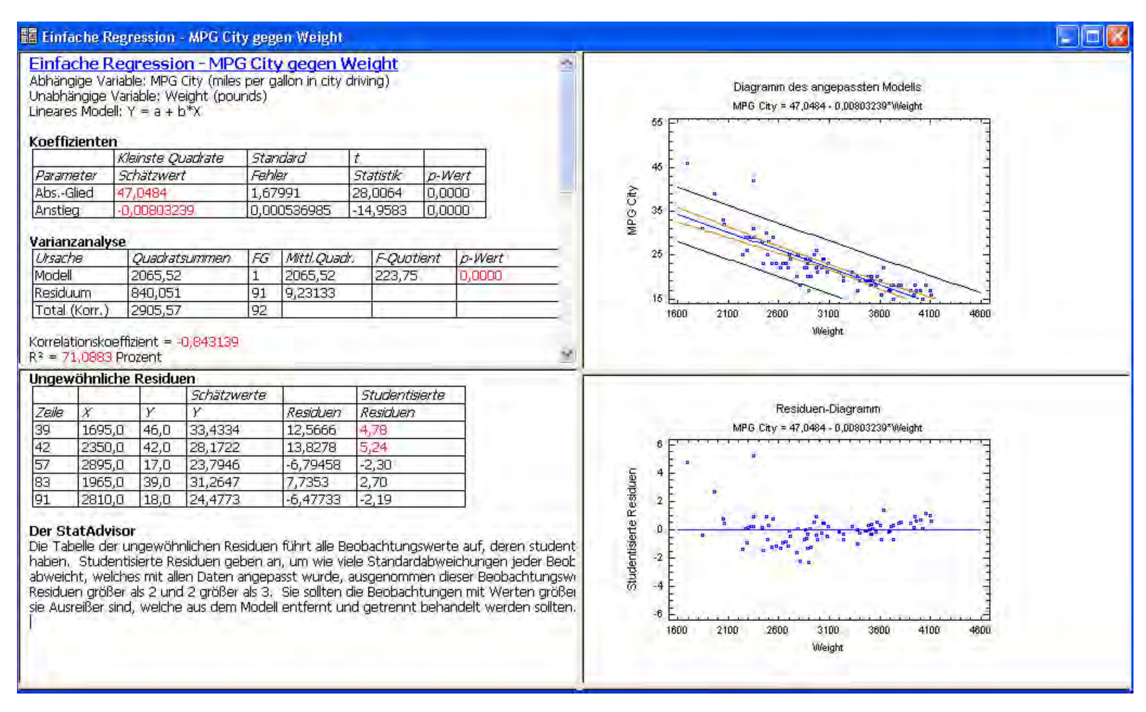
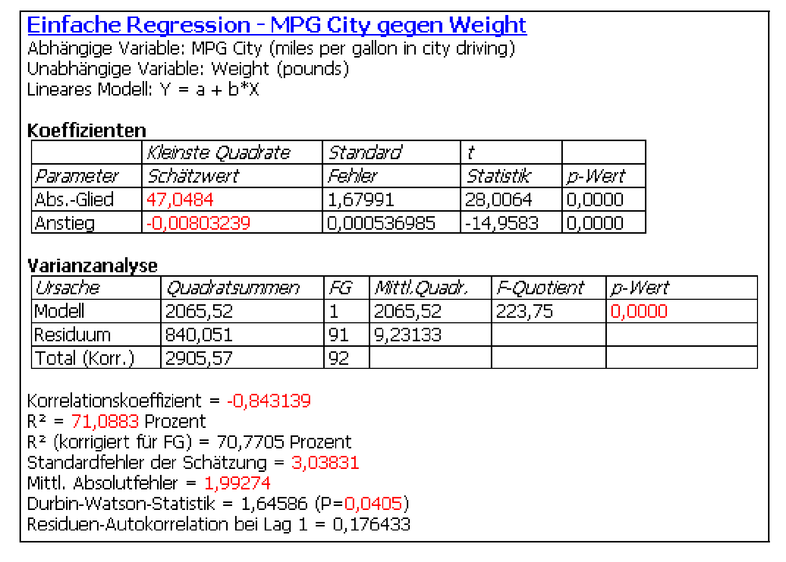


Abbildung 13.6: Analyse-Fenster der Einfachen Regression

Die *Zusammenfassung der Analyse* im oberen linken Ergebnisfenster fasst die Anpassung zusammen:



Einfache Regression - MPG City gegen Weight
Abhängige Variable: MPG City (miles per gallon in city driving)
Unabhängige Variable: Weight (pounds)
Lineares Modell: Y = a + b*X

Koeffizienten

	Kleinste Quadrate	Standard	t	
Parameter	Schätzwert	Fehler	Statistik	p-Wert
Abs.-Glied	47,0484	1,67991	28,0064	0,0000
Anstieg	-0,00803239	0,000536985	-14,9583	0,0000

Varianzanalyse

Ursache	Quadratsummen	FG	Mittl.Quadr.	F-Quotient	p-Wert
Modell	2065,52	1	2065,52	223,75	0,0000
Residuum	840,051	91	9,23133		
Total (Korr.)	2905,57	92			

Korrelationskoeffizient = -0,843139
R² = 71,0883 Prozent
R² (korrigiert für FG) = 70,7705 Prozent
Standardfehler der Schätzung = 3,03831
Mittl. Absolutfehler = 1,99274
Durbin-Watson-Statistik = 1,64586 (P=0,0405)
Residuen-Autokorrelation bei Lag 1 = 0,176433

Abbildung 13.7: Analyse-Zusammenfassung für die Einfache Regression

Von den vielen Statistiken der oberen Tabelle sind die folgenden besonders wichtig:

1. **Koeffizienten**: geschätzte Modellkoeffizienten. Das angepasste Modell, das man zur Vorhersage verwenden könnte, ist folgendes:

$$MPG\ City = 47{,}0484 - 0{,}00803239\,Weight$$

2. **R²**: prozentualer Anteil an der Variabilität von Y, der durch das Modell erklärt wird. In unserem Fall erklärt eine lineare Regression mit *Weight* ungefähr 71,1 % der Variabilität von *MPG City*.

3. ***p*-Wert des Modells** : testet die Null-Hypothese, dass das angepasste Modell nicht besser ist als das triviale Modell, welches *Weight* nicht enthält. Ein *p*-Wert unter 0,05 (wie in unserem Beispiel) bedeutet, dass *Weight* eine sinnvolle Einflussgröße für *MPG City* darstellt.

Das Diagramm im oberen rechten Ergebnisfenster zeigt das angepasste Modell:

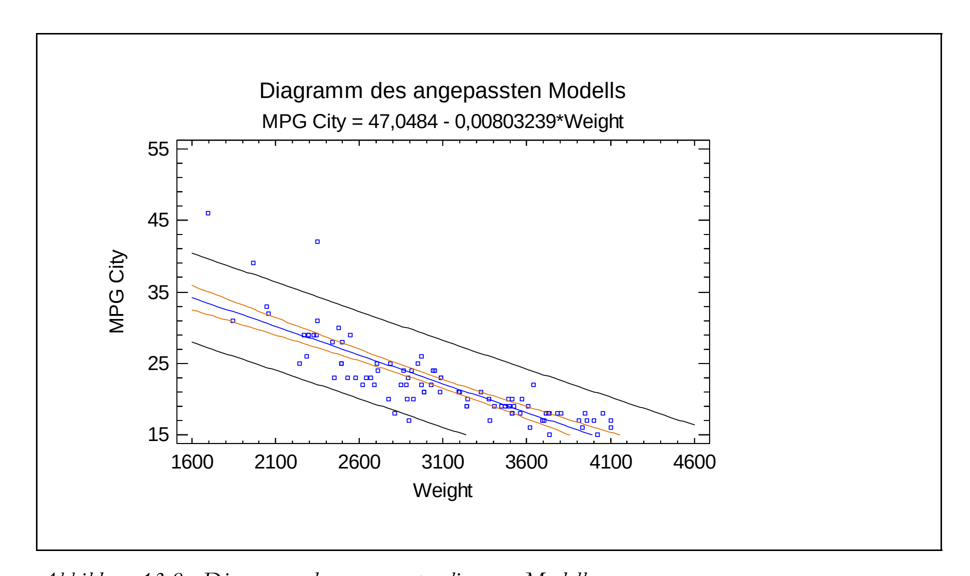


Abbildung 13.8: Diagramm des angepassten linearen Modells

Das Diagramm zeigt die Kleinste-Quadrate-Regressionsgerade und zwei Paare von Grenzen. Die inneren Grenzen bilden 95%-Konfidenzintervalle für den Mittelwert von Y bei vorgegebenem X. Diese zeigen an, wie gut die Lage der Geraden geschätzt wurde, unter der Voraussetzung, dass die Beziehung linear ist. Je größer die Stichprobe, umso enger sind diese Grenzen. Die äußeren Grenzen sind 95%-Vorhersagegrenzen für eine neue Beobachtung.

Schätzungsweise 95 % von zusätzlichen Beobachtungen, ähnlich denen in unserer Stichprobe, würden innerhalb dieser Grenzen liegen.

Es ist beachtenswert, dass drei Beobachtungen für kleine Werte von *Weight* ziemlich weit außerhalb der 95%-Vorhersagegrenzen liegen. Das kann ein Indikator sein entweder für Ausreißer oder für einen Modellfehler, der durch die Nichtlinearität der tatsächlichen Beziehung zwischen *MPG City* und *Weight* erklärbar ist.

13.3 Anpassen eines nichtlinearen Modells

Die Prozedur *Einfache Regression* bietet die Möglichkeit, ein breites Spektrum nichtlinearer Modelle anzupassen. Um diese relative Verbesserung, die für verschiedene Modelle verfügbar ist, zugänglich zu machen, wählen Sie *Vergleich der alternativen Modelle* in der Dialogbox *Tabellen und Grafiken*. Dann werden alle denkbaren Modelle angepasst und in absteigender Reihenfolge nach R^2 geordnet aufgelistet:

Vergleich der alternativen Modelle

Modell	*Korrelation*	*R²*
S-Kurven - Modell	0,9016	81,29%
Reziprok in Y, Quadratwurzel-X	0,8995	80,92%
Reziprok-Y, logarithmisch in X	0,8995	80,90%
Quadratwurzel-Y, reziprok-X	0,8988	80,78%
Multiplikativ	-0,8981	80,65%
Reziprok-Y	0,8969	80,44%
Logarithmisch in Y, Quadratwurzel-X	-0,8919	79,54%
Doppelt reziprok	-0,8896	79,14%
Reziprok-X	0,8888	79,00%
Quadratwurzel-Y, logarithmisch in X	-0,8879	78,83%
Reziprok-Y, quadratisch in X	0,8852	78,35%
Exponentiell	-0,8833	78,03%
Doppelte Quadratwurzel	-0,8784	77,16%
Logarithmisch-X	-0,8705	75,78%
Quadratwurzel-Y	-0,8668	75,14%
Logarithmisch in Y, quadratisch in X	-0,8611	74,15%
Quadratwurzel-X	-0,8577	73,56%
Quadratisch in Y, reziprok in X	0,8472	71,77%
Linear	-0,8431	71,09%
Quadratwurzel-Y, quadratisch in X	-0,8393	70,44%
Quadratisch in Y, logarithmisch in X	-0,8146	66,35%
Quadratisch in X	-0,8106	65,71%
Quadratisch in Y, Quadratwurzel-X	-0,7957	63,31%
Y-Quadrat	-0,7758	60,18%
Doppelt quadratisch	-0,7346	53,96%
Logistisch	<keine Anp.>	
Log-Probit	<keine Anp.>	

Abbildung 13.9: Vergleich der alternativen Modelle

Die Modelle ganz oben in der Liste erklären den größten prozentualen Anteil an der Variation der Zielgröße. R^2 ist nur eines der Kriterien, das man als Hilfe zum Auswählen eines Modells verwenden kann. Modelle mit etwas niedrigerem R^2 als das oberste Modell wären eventuell vorzuziehen, sofern sie im Kontext der Daten mehr Sinn machen.

In unserem Beispiel ist das *Reziproke-Y-Modell*, das ziemlich weit oben aufgelistet ist, ein attraktives Modell. Es hat die Form:

$$\frac{1}{MPGCity} = \beta_0 + \beta_1 Weight + \varepsilon$$

Damit wird der Kehrwert von *MPG City* (also Gallonen pro Meile) als lineare Funktion des Gewichtes ausgedrückt. Es ist nicht ungewöhnlich, dass Transformationen von Y, X oder beiden zu besseren Modellen führen.
Um das *Reziproke-Y-Modell* anzupassen, drücken Sie die Schaltfläche *Analyse-Optionen* und wählen in dieser Dialogbox *Reziprok-Y* aus. Das Ergebnis der Modellanpassung sieht folgendermaßen aus:

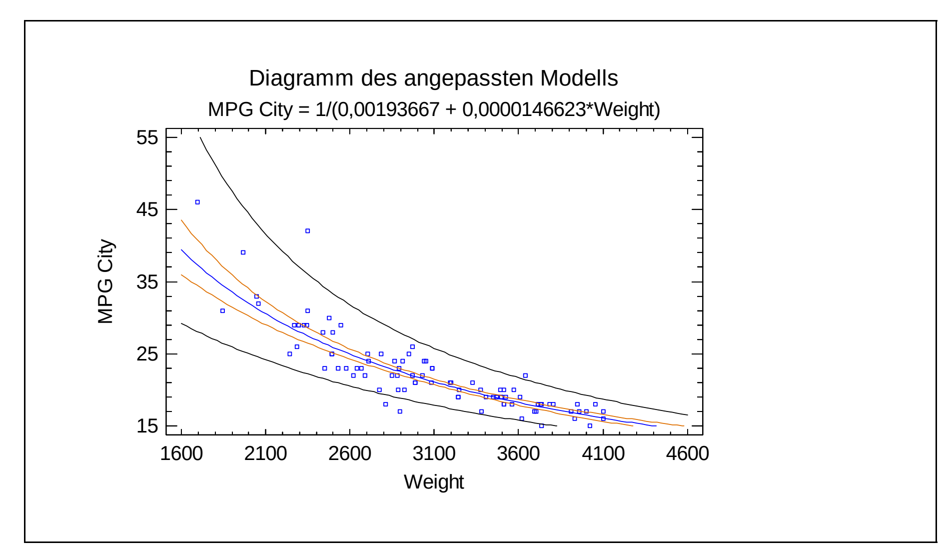


Abbildung 13.10: Angepasstes Reziprok-Y-Modell

Während dieses Modell linear im Kehrwert von *MPG City* ist, ist es nichtlinear in der originalen Metrik. Beachten Sie auch, dass die Vorhersagegrenzen sich verbreitern, wenn die Vorhersagewerte größer werden. Im Kontext der Daten macht das Sinn, da es impliziert, dass unter den leichteren Autos eine größere Variabilität herrscht als unter den schwereren.

13.4 Prüfen der Residuen

Sobald ein vernünftiges Modell angepasst worden ist, sollten die Residuen untersucht werden. Im Allgemeinen kann man sich ein Residuum als die Differenz zwischen dem beobachteten Wert von Y und dem Modellwert vorstellen:

Residuum = beobachtetes Y – vorhergesagtes Y

Die *Einfache Regression* erzeugt automatisch Diagramme der Residuen gegen die X-Variable:

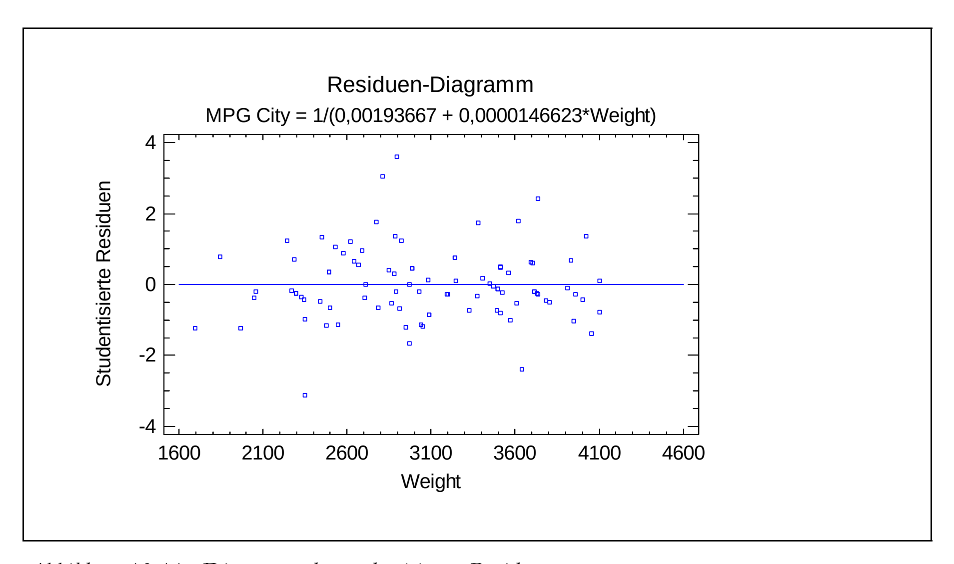


Abbildung 13.11: Diagramm der studentisierten Residuen

Mittels *Ergebnisfenster-Optionen* kann man auswählen, ob man die einfachen Residuen oder die studentisierten Residuen zeichnen möchte. Studentisierte Residuen sind die einfachen Residuen dividiert durch ihre geschätzten Standardfehler. Ein studentisiertes Residuum drückt daher aus, um wie viele Standardfehler die Datenwerte vom angepassten Modell entfernt sind.

STATGRAPHICS Centurion XVI verwendet in der Regel studentisierte Residuen mit Ausschluss. Residuen mit Ausschluss werden berechnet, indem jeweils eine der Beobachtungen ausgeschlossen und mit allen anderen das Modell geschätzt wird. Dann wird der Abstand der ausgeschlossenen Beobachtung vom Modellwert als Vielfaches der Standardabweichung bestimmt. Diese Vorgehensweise verhindert, dass Ausreißer einen großen Einfluss auf das berechnete Modell ausüben, mit dem ihr Residuum bestimmt wird. Die Auswahl *Ungewöhnliche Residuen* in der Dialogbox *Tabellen und Grafiken* listet alle studentisierten Residuen auf, deren Absolutwert größer als 2 ist.

Ungewöhnliche Residuen

			Schätzwerte		Studentisierte
Zeile	X	Y	Y	Residuen	Residuen
39	1695,0	46,0	33,4334	12,5666	
42	2350,0	42,0	28,1722	13,8278	
57	2895,0	17,0	23,7946	-6,79458	-2,30
83	1965,0	39,0	31,2647	7,7353	2,70
91	2810,0	18,0	24,4773	-6,47733	-2,19

Abbildung 13.12: Tabelle der ungewöhnlichen Residuen

Studentisierte Residuen, die größer als 3 sind, wie in Zeile 57, sind potentielle Ausreißer, die anscheinend nicht zu den anderen Daten gehören. Zeile 57 steht für einen Mazda RX-7, für den ein *MPG City*-Wert von nur 17 aufgezeichnet wurde, obwohl das Modell den Wert 22,5 vorhersagt. Im nächsten Abschnitt werden dem Modell weitere Variablen hinzugefügt, was die Vorhersagefähigkeit für solche Sportwagen verbessern kann; Zeile 57 wird daher nicht von der Anpassung ausgeschlossen, obwohl dieser Fall besonders zu beachten ist.

13.5 Mehrfache Regression

Um das Modell zu verbessern, ist es sinnvoll, weitere Einflussvariablen hinzuzufügen. Am einfachsten ist das mit der Prozedur *Mehrfache Regression,* die man im Hauptmenü auf folgende Weise findet:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beziehungen – Mehrere Faktoren – Mehrfache Regression.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Verbessern – Regressionsanalyse – Mehrere Faktoren– Mehrfache Regression.*

Die Dateneingabe-Dialogbox sieht folgendermaßen aus:

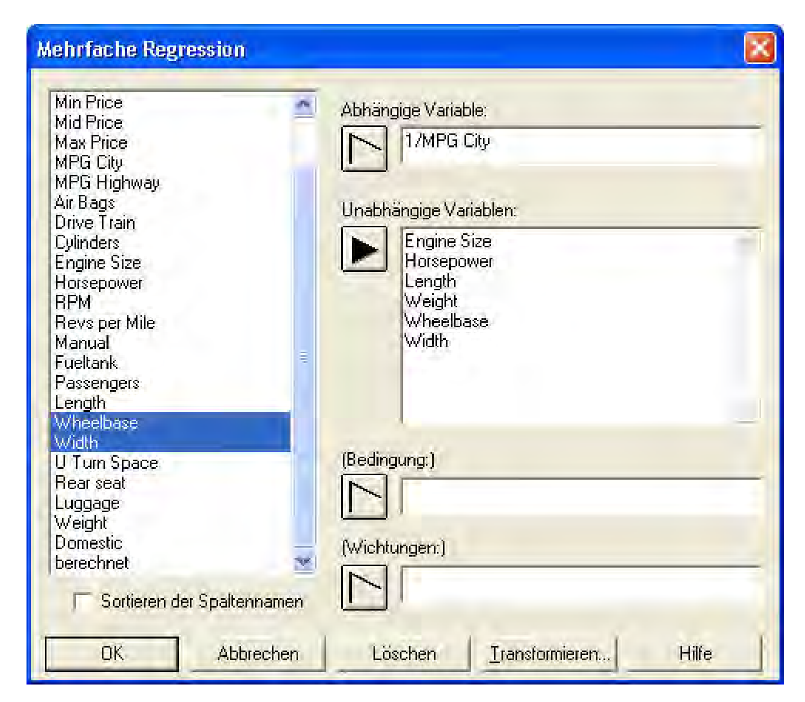


Abbildung 13.13: Dateneingabe-Dialogbox der Mehrfachen Regression

Zu Beginn werden alle sechs Einflussgrößen, die bereits früher in der Prozedur *Analyse mehrerer Variablen* diskutiert wurden, als unabhängige Variablen eingegeben. Als abhängige Variable wird der Kehrwert von *MPG City*, der den Gallonen pro Meile entspricht, verwendet. Anschließend werden das *Optionen*-Menü und die Dialogbox *Tabellen und Grafiken* geöffnet. Die aus den Einstellungen resultierende Zusammenfassung der Analyse wird nachfolgend angezeigt:

Mehrfache Regression - 1/MPG City
Abhängige Variable: 1/MPG City
Unabhängige Variable:
Engine Size (liters)
Horsepower (maximum)
Length (inches)
Weight (pounds)
Wheelbase (inches)
Width (inches)

		Standard	T	
Parameter	Schätzwert	Fehler	Statistik	p-Wert
KONSTANTE	0,0155897	0,0177088	0,880334	0,3811
Engine Size	0,00072849	0,000980504	0,742974	0,4595
Horsepower	0,0000132632	0,000014911	0,889485	0,3762
Length	-0,000101355	0,0000608857	-1,66468	0,0996
Weight	0,0000149727	0,00000242804	6,1666	0,0000
Wheelbase	-0,000148122	0,000163073	-0,908321	0,3662
Width	0,000223526	0,00028967	0,771658	0,4424

Varianzanalyse

Ursache	Quadratsummen	FG	Mittl.Quadr.	F-Quotient	p-Wert
Modell	0,00705967	6	0,00117661	67,64	0,0000
Residuum	0,001496	86	0,0000173954		
Total (Korr.)	0,00855567	92			

R^2 = 82,5145 Prozent
R^2 (korrigiert für FG) = 81,2946 Prozent
Standardfehler der Schätzung = 0,00417077
Mittl. Absolutfehler = 0,00304978
Durbin-Watson-Statistik = 1,6264 (P=0,0306)
Residuen-Autokorrelation bei Lag 1 = 0,186005

Der StatAdvisor
Hier werden die Ergebnisse der Anpassung eines multiplen linearen Modells zur Beschreibung des (funktionalen) Zusammenhangs zwischen 1/MPG City und 6 unabhängigen Variablen angezeigt. Die Gleichung des angepassten Modells ist die folgende:

1/MPG City = 0,0155897 + 0,00072849*Engine Size + 0,0000132632*Horsepower - 0,000101355*Length + 0,0000149727*Weight - 0,000148122*Wheelbase + 0,000223526*Width

Da der p-Wert in der ANOVA-Tabelle kleiner 0,05 ist, existiert eine statistisch signifikante Beziehung zwischen den Variablen bei einem Konfidenzniveau von 95,0%.

Abbildung 13.14: Analyse-Zusammenfassung für die Mehrfache Regression mit sechs Einflussgrößen

Beachten Sie, dass die R^2-Statistik auf 82,5 % angestiegen ist. Jedoch ist das Modell viel zu komplex. Oben in der Ausgabe ist eine Spalte mit *p*-Werten enthalten. Diese *p*-Werte sind Ergebnisse des Tests mit der Hypothese, dass der zur ausgewählten Variable gehörige Koeffizient gleich 0 ist, sofern alle Variablen im Modell verbleiben. Ein *p*-Wert größer als 0,05 bedeutet, dass die betreffende Variable beim Vorhandensein aller anderen Variablen keinen signifikanten Beitrag zum Modell liefert.

Mit Ausnahme von *Weight* haben alle anderen Einflussvariablen *p*-Werte größer als 0,05. Das bedeutet, dass mindestens eine dieser Einflussgrößen aus dem Modell entfernt werden könnte, ohne dass die Modellsignifikanz darunter leiden würde.

ANMERKUNG: An dieser Stelle wäre es falsch anzunehmen, alle fünf Einflussgrößen mit *p*-Werten über 0,05 könnten entfernt werden. Wegen der hohen Multikollinearität in

den Daten können sich die *p*-Werte dramatisch ändern, wenn auch nur eine Variable aus dem Modell entfernt wird.

Eine sinnvolle Vorgehensweise für die Vereinfachung des Modells ist das Ausführen einer schrittweisen Regression. In einer schrittweisen Regression werden die Variablen eine nach der anderen dem Modell hinzugefügt oder daraus entfernt, mit dem Ziel, ein Modell zu erreichen, welches nur signifikante Einflussgrößen enthält und keine sinnvollen Variablen ausschließt. Die schrittweise Regression ist als Option in der Dialogbox *Analyse-Optionen* verfügbar:

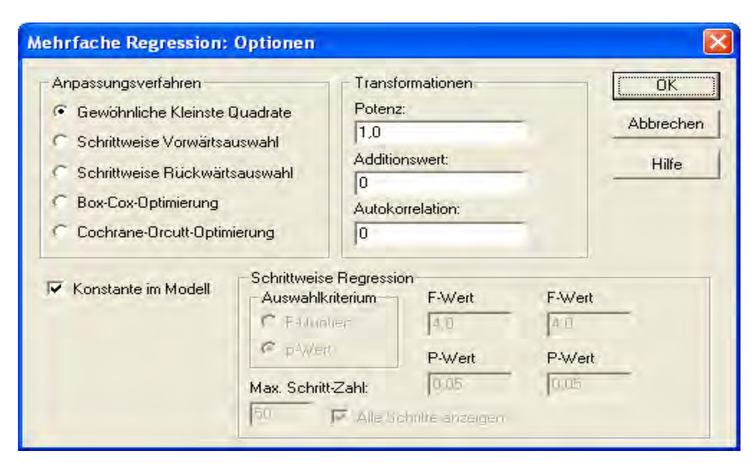


Abbildung 13.15: Analyse-Optionen für die Mehrfache Regression

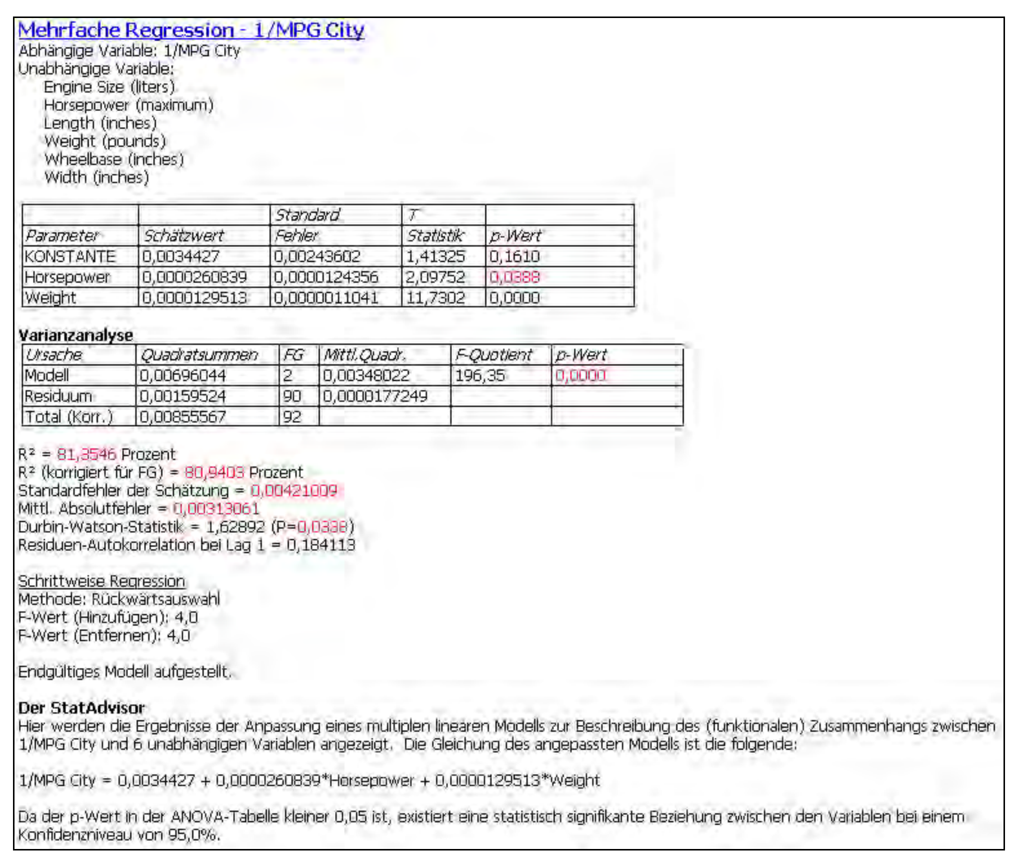
Es werden zwei Optionen für die schrittweise Regression angeboten:

1. *Vorwärtsauswahl* – beginnt mit einem Modell, welches nur das konstante Glied enthält, und bringt eine Variable nach der anderen in das Modell hinein, falls sie die Modellsignifikanz verbessert.

2. *Rückwärtsauswahl* – beginnt mit einem Modell, welches alle Variablen enthält, und entfernt diese eine nach der anderen, solange, bis alle verbleibenden Variablen signifikant sind.

In beiden Verfahren können bereits entfernte Variablen in einem späteren Schritt wieder hinzukommen, wenn sie dann als sinnvolle Einflussgrößen erscheinen, bzw. zunächst hinzugefügte Variablen können später wieder entfernt werden, wenn sie nicht mehr signifikant sind.

Das Durchführen einer Rückwärtsauswahl ergibt folgendes Modell:


Mehrfache Regression - 1/MPG City
Abhängige Variable: 1/MPG City
Unabhängige Variable:
Engine Size (liters)
Horsepower (maximum)
Length (inches)
Weight (pounds)
Wheelbase (inches)
Width (inches)

		Standard	T	
Parameter	Schätzwert	Fehler	Statistik	p-Wert
KONSTANTE	0,0034427	0,00243602	1,41325	0,1610
Horsepower	0,0000260839	0,0000124356	2,09752	0,0388
Weight	0,0000129513	0,0000011041	11,7302	0,0000

Varianzanalyse

Ursache	Quadratsummen	FG	Mittl.Quadr.	F-Quotient	p-Wert
Modell	0,00696044	2	0,00348022	196,35	0,0000
Residuum	0,00159524	90	0,0000177249		
Total (Korr.)	0,00855567	92			

R² = 81,3546 Prozent
R² (korrigiert für FG) = 80,9403 Prozent
Standardfehler der Schätzung = 0,00421009
Mittl. Absolutfehler = 0,00313061
Durbin-Watson-Statistik = 1,62892 (P=0,0338)
Residuen-Autokorrelation bei Lag 1 = 0,184113

Schrittweise Regression
Methode: Rückwärtsauswahl
F-Wert (Hinzufügen): 4,0
F-Wert (Entfernen): 4,0

Endgültiges Modell aufgestellt.

Der StatAdvisor
Hier werden die Ergebnisse der Anpassung eines multiplen linearen Modells zur Beschreibung des (funktionalen) Zusammenhangs zwischen 1/MPG City und 6 unabhängigen Variablen angezeigt. Die Gleichung des angepassten Modells ist die folgende:

1/MPG City = 0,0034427 + 0,0000260839*Horsepower + 0,0000129513*Weight

Da der p-Wert in der ANOVA-Tabelle kleiner 0,05 ist, existiert eine statistisch signifikante Beziehung zwischen den Variablen bei einem Konfidenzniveau von 95,0%.

Abbildung 13.16: Analyse-Zusammenfassung für die Mehrfache Regression mit Rückwärtsauswahl

Nur zwei Variablen bleiben im Modell: *Horsepower* und *Weight*. Beide Variablen haben *p*-Werte unter 0,05.

Sobald eine mathematische Gleichung aufgestellt ist, ist es aufschlussreich, diese Gleichung grafisch darzustellen. Da das Modell zwei Einflussgrößen enthält, repräsentiert die Modellgleichung eine Fläche im dreidimensionalen Raum, welche in der Regel als *Wirkungsfläche* (oder *(Ober-)Fläche)* bezeichnet wird. In unserem Fall entspricht die angepasste Gleichung einer Ebene, da *Horsepower* und *Weight* in linearer Form im Modell enthalten sind.

Zur grafischen Darstellung des Modells kann man

die Prozedur *Flächen- und Konturen-Diagramme* verwenden, indem Sie die darzustellende Funktion hineinkopieren und Ihre eigenen Beschriftungen und Skalierungen folgendermaßen festlegen:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Diagramme – Flächen- und Konturen-Diagramme*.

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Werkzeuge – Flächen- und Konturen-Diagramme*.

In der Dateneingabe-Dialogbox geben Sie das Modell ein, wobei Sie die beiden Einflussgrößen mit X und Y bezeichnen. Der einfachste Weg wäre, die bei der *Mehrfachen Regression* erzeugte Gleichung zu kopieren und hier wieder einzufügen und anschließend *Horsepower* in X und *Weight* in Y zu ändern:

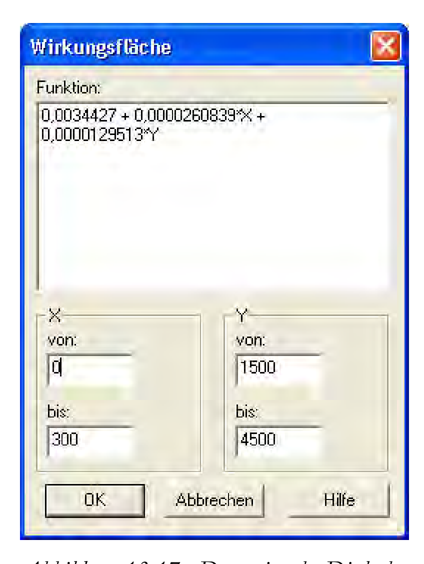


Abbildung 13.17: Dateneingabe-Dialogbox für Flächen- und Konturen-Diagramme

Die Skalierung für X und Y sollte ebenfalls geändert werden, so dass sie für die zur Modellanpassung verwendeten Daten geeignet ist.

Nach dem Anklicken von *OK* wird zunächst die Dialogbox *Tabellen und Grafiken* angezeigt. Anschließend wird das Flächen-Diagramm erstellt. Die anfängliche Darstellung hat das Aussehen einer Drahtgitteroberfläche:

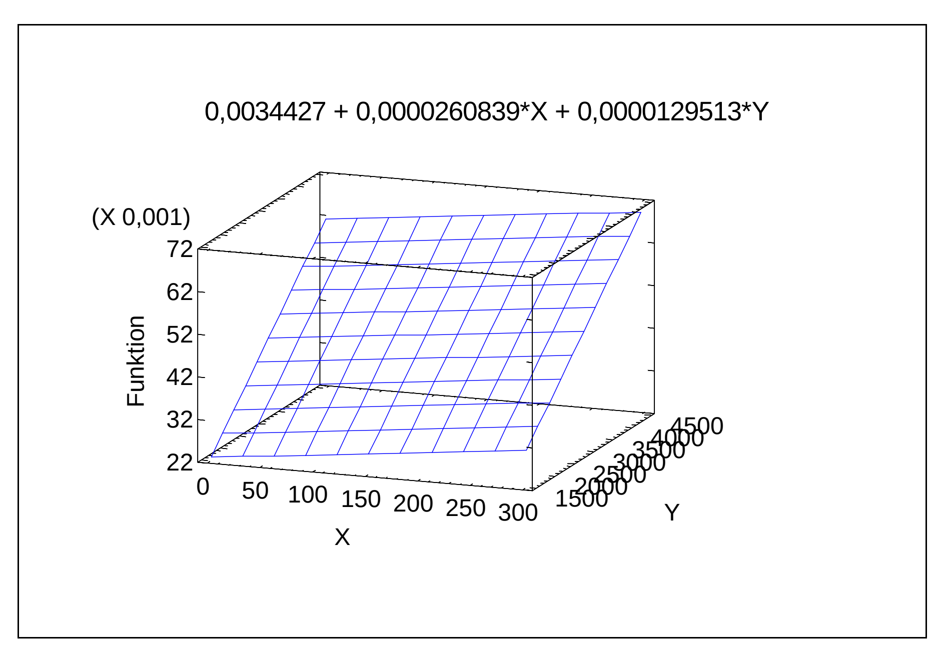


Abbildung 13.18: Flächen-Diagramm mit Standardbeschriftungen und -Skalierung

Man kann die Darstellung stark verändern durch:

- Auswahl der *Diagramm-Optionen* auf der Analyse-Symbolleiste und Verändern der Beschriftungen und Skalierungen in den Registern *Titel, x-Achse*, *y-Achse* und *z-Achse*. Speziell:
  - Ändern Sie den x-Achsen-Titel in *Horsepower*.
  - Ändern Sie den y-Achsen-Titel in *Weight*.
  - Ändern Sie die y-Achsenskalierung: von 1500 bis 4500 mit dem Abstand 1000.
  - Ändern Sie den z-Achsen-Titel in *1/MPG City*.
  - Ändern Sie den Legenden-Titel in *1/MPG City*.
- Auswahl der *Ergebnisfenster-Optionen* und Verändern des anzuzeigenden Diagrammtyps:

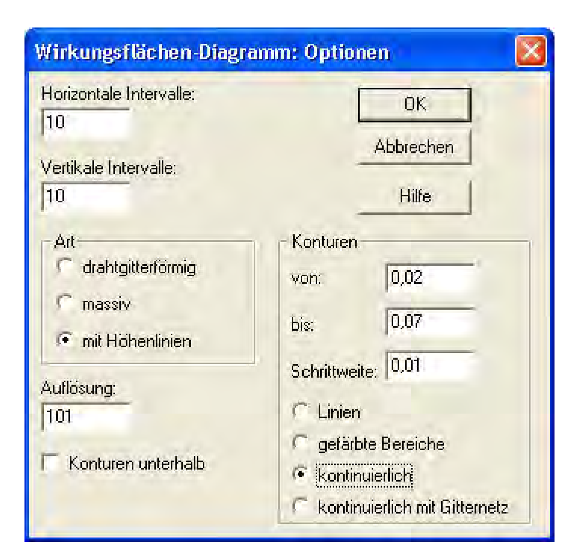


Abbildung 13.19: Ergebnisfenster-Optionen für Flächen-Diagramme

In der obigen Dialogbox wurde das Gruppenfeld *Art* auf *mit Höhenlinien* gesetzt und *Konturen* auf *kontinuierlich*. Das fertige Diagramm sieht so aus:

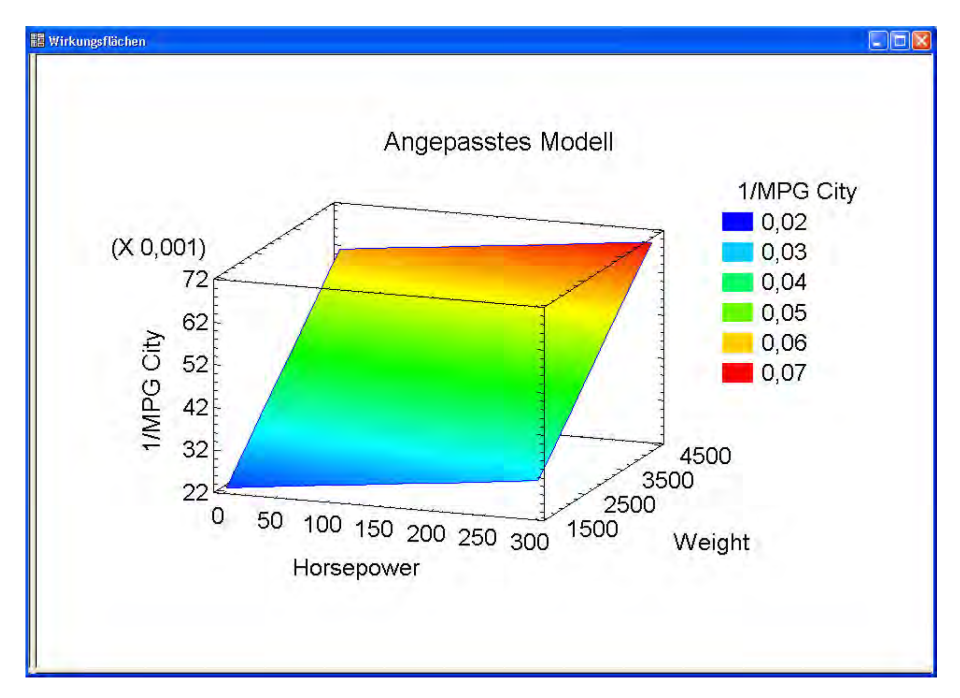


Abbildung 13.20: Diagramm des angepassten Modells

Die Autos, die den meisten Kraftstoff verbrauchen, sind in der hinteren rechten Ecke zu finden: große Autos mit großen Motoren.

Kapitel

14

Tutorial 5: Analysieren attributiver Daten

Häufigkeitstabellen, Kontingenztafeln und eine Pareto-Analyse.

Jedes der ersten vier Tutorials beschäftigte sich mit messbaren Variablen, bei denen die Beobachtungen durch Zahlen auf einer kontinuierlichen Skala repräsentiert wurden. In diesem Tutorial werden eine Reihe attributiver Daten untersucht, bei denen jede Beobachtung anstelle eines Messwertes eine bestimmte Kategorie repräsentiert, in welche die Variable unterteilt ist.

Als Beispiel betrachten wir die Daten, die in der Datei *defects.sgd* enthalten sind. Ein Teil dieser Datei wird hier angezeigt:

Defect	*Facility*
Misaligned	Virginia
Contaminated	Texas
Contaminated	Virginia
Contaminated	Texas
Missing parts	Texas
Misaligned	Virginia
Contaminated	Texas
Leaking	Texas
Damaged	Virginia
Contaminated	Texas

Die Datei enthält n = 120 Zeilen, jede entspricht einem Fehler, der an einem produzierten Gegenstand beobachtet wurde. Die Datei zeigt die Art des Fehlers an (Spalte *Defect*) und die Anlage, in welcher der Gegenstand produziert wurde (Spalte *Facility*).

14.1 Zusammenfassungen für attributive Daten

Wir ignorieren für einen Moment die Anlage, in welcher jeder der Gegenstände produziert wurde, und interessieren uns zunächst für die Zusammenfassung der Fehlertypen:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beschreiben – Kategoriale Daten – Tabellierung*.

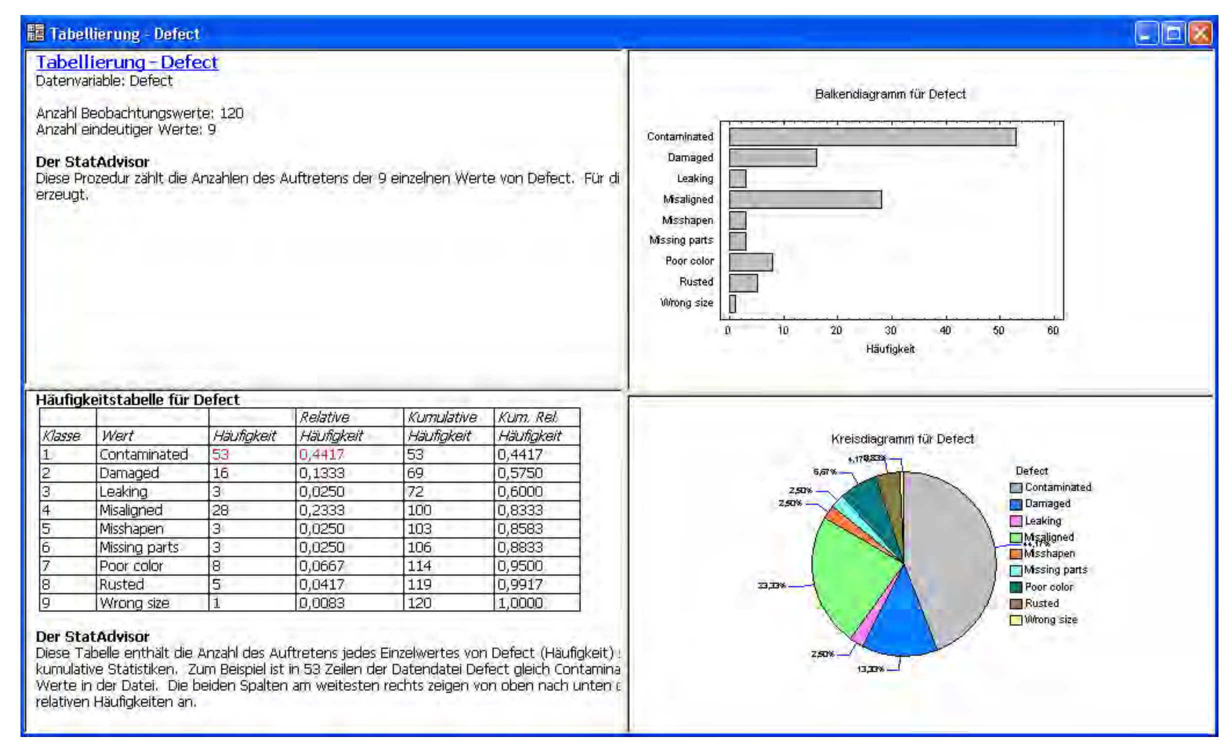
2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Attributive Daten – Ein Faktor – Tabellierung*.

Die Dateneingabe-Dialogbox benötigt eine einzelne Spalte, welche die attributiven Daten enthält:



Abbildung 14.1: Dateneingabe-Dialogbox für die Tabellierung

Die Prozedur begutachtet die Spalte und identifiziert die eindeutigen Werte. Die Dialogbox *Tabellen und Grafiken* wird geöffnet und anschließend wird ein Analyse-Fenster ähnlich diesem hier erstellt:


Tabellierung - Defect

Tabellierung - Defect
Datenvariable: Defect

Anzahl Beobachtungswerte: 120
Anzahl eindeutiger Werte: 9

Der StatAdvisor
Diese Prozedur zählt die Anzahlen des Auftretens der 9 einzelnen Werte von Defect. Für di
erzeugt.

Häufigkeitstabelle für Defect

			Relative	Kumulative	Kum. Rel.
Klasse	Wert	Häufigkeit	Häufigkeit	Häufigkeit	Häufigkeit
1	Contaminated	53	0,4417	53	0,4417
2	Damaged	16	0,1333	69	0,5750
3	Leaking	3	0,0250	72	0,6000
4	Misaligned	28	0,2333	100	0,8333
5	Misshapen	3	0,0250	103	0,8583
6	Missing parts	3	0,0250	106	0,8833
7	Poor color	8	0,0667	114	0,9500
8	Rusted	5	0,0417	119	0,9917
9	Wrong size	1	0,0083	120	1,0000

Der StatAdvisor
Diese Tabelle enthält die Anzahl des Auftretens jedes Einzelwertes von Defect (Häufigkeit)
kumulative Statistiken. Zum Beispiel ist in 53 Zeilen der Datendatei Defect gleich Contamina
Werte in der Datei. Die beiden Spalten am weitesten rechts zeigen von oben nach unten
relativen Häufigkeiten an.

Abbildung 14.2: Analyse-Fenster der Tabellierung

Das obere linke Ergebnisfenster zeigt an, dass neun eindeutige Werte in den n = 120 Zeilen gefunden wurden. Das Balkendiagramm und das Kreisdiagramm rechts veranschaulichen die beobachteten Häufigkeiten für alle Fehlerarten, die auch als Tabelle im unteren linken Ergebnisfenster enthalten sind. Der am häufigsten vorkommende Fehler ist *Contaminated* mit einem Anteil von 44 % an allen Fehlern.

14.2 Pareto-Analyse

Die Prozedur *Tabellierung* (Häufigkeitstabellen) sortiert die Fehlerarten alphabetisch. Um sie nach der Häufigkeit zu sortieren, kann stattdessen die Prozedur *Pareto-Analyse* verwendet werden. Diese ist folgendermaßen zu finden:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *SPC – Qualitätsbeurteilung – Pareto-Analyse.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Attributive Daten – Ein Faktor – Pareto-Analyse.*

Die Dateneingabe-Dialogbox sollte folgendermaßen ausgefüllt werden:

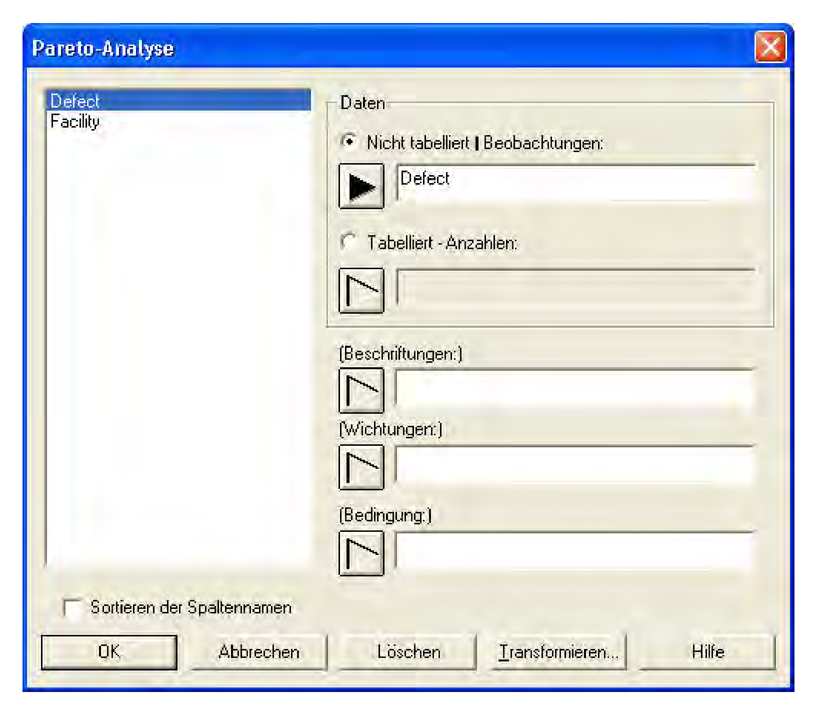


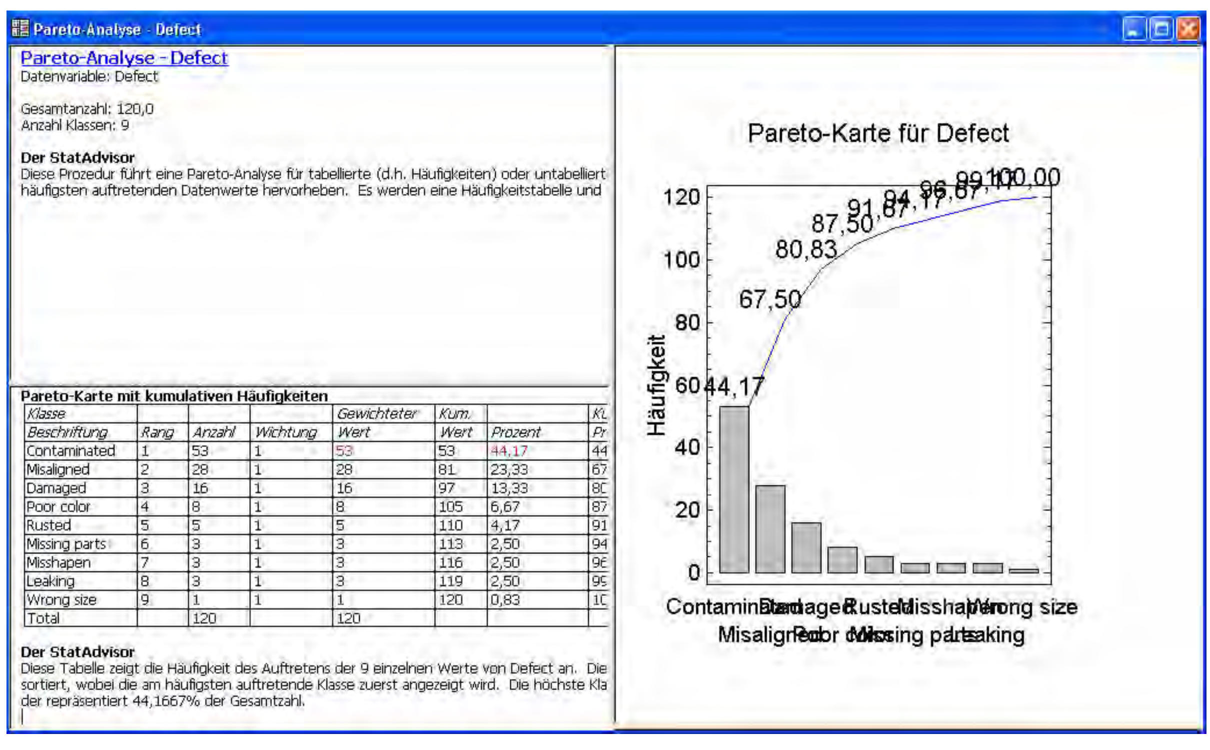
Abbildung 14.3: Dateneingabe-Dialogbox der Pareto-Analyse

Die *Pareto-Analyse* akzeptiert Daten in zwei Varianten:

1. Nichttabellierte Daten, deren Häufigkeiten noch ausgezählt werden müssen, so wie in unserem Beispiel.

2. Ausgezählte Häufigkeiten je Fehlergruppe. Dies ist anwendbar, wenn man zwei Spalten hat, eine, die den Fehlertyp identifiziert, und eine zweite, welche die Anzahl des Auftretens für diesen Fehler enthält.

Das Analyse-Fenster enthält sowohl eine Häufigkeitstabelle als auch ein Pareto-Diagramm:



Klasse Beschriftung	Rang	Anzahl	Wichtung	Gewichteter Wert	Kum. Wert	Prozent	Kl. Pr
Contaminated	1	53	1	53	53	44,17	44
Misaligned	2	28	1	28	81	23,33	67
Damaged	3	16	1	16	97	13,33	8C
Poor color	4	8	1	8	105	6,67	87
Rusted	5	5	1	5	110	4,17	91
Missing parts	6	3	1	3	113	2,50	94
Misshapen	7	3	1	3	116	2,50	9E
Leaking	8	3	1	3	119	2,50	99
Wrong size	9	1	1	1	120	0,83	1C
Total		120		120			

Abbildung 14.4: Analyse-Fenster der Pareto-Analyse

Von besonderem Interesse ist das Pareto-Diagramm auf der rechten Seite, welches die Anzahlen pro Fehlerart von der häufigsten bis zu der am wenigsten häufigen darstellt. Anfangs überlappen leider die Beschriftungen wegen ihrer Länge und Anzahl. Das kann in Ordnung gebracht werden, indem man:

1. mit der Maus doppelt in die Grafik klickt, um das Ergebnisfenster innerhalb des Analyse-Fensters zu maximieren.

2. die Schaltfläche *Diagramm-Optionen* auf der Analyse-Symbolleiste anklickt, das Register *x-Achse* wählt und das Kontrollkästchen *Achsenbeschriftung drehen* aktiviert.

3. Nach dem Verlassen der Dialogbox *Diagramm-Optionen* kann es passieren, dass die Beschriftungen nicht komplett auf den Bildschirm passen. Ist dies der Fall, so kann das Diagramm bei in Diagramm-Mitte platzierter und gedrückter Maustaste in eine andere Position gezogen werden. Oder man zieht die x-Achse nach oben, um die Höhe der vertikalen Achse zu reduzieren.

Nach Fertigstellung sollte das Pareto-Diagramm wie folgt aussehen:

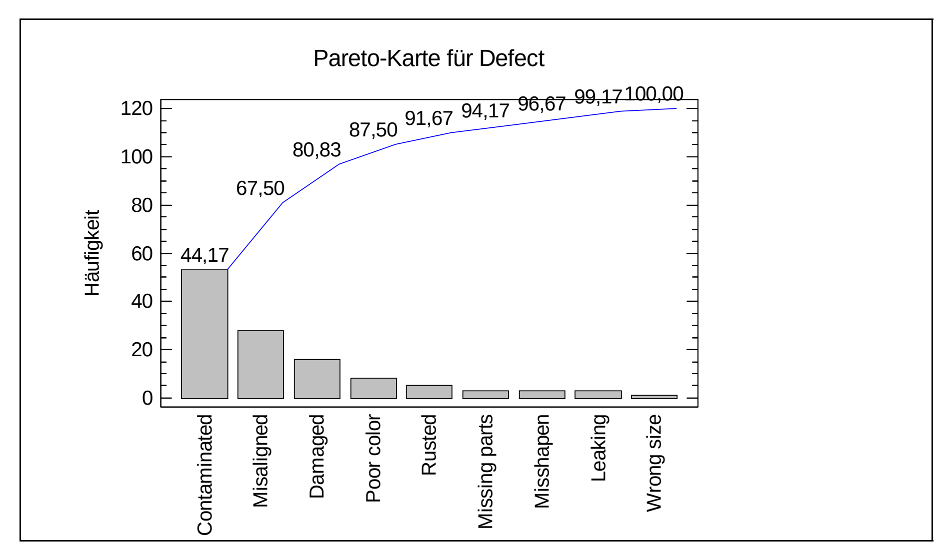


Abbildung 14.5: Bearbeitetes Pareto-Diagramm

Die vertikalen Säulen im Pareto-Diagramm haben eine Höhe proportional zur absoluten Fehlerhäufigkeit. Die darüber verlaufende Linie stellt die kumulierte Summe der Fehlerzahlen von links nach rechts dar. Über jeder Säule wird die von links kumulierte relative Fehlerhäufigkeit in Prozent angezeigt.

Das grundlegende Pareto-Prinzip besagt, dass die große Mehrheit der Fehler auf eine kleine Zahl möglicher Ursachen zurückzuführen ist. In unserem Fall sind die drei häufigsten Fehlerarten für mehr als 80 % aller auftretenden Fehler verantwortlich.

14.3 Kreuztabellen

Die Datei *defects.sgd* enthält zudem eine Spalte zur Identifizierung der Anlage (*facility*), in der die Gegenstände produziert wurden. Zur Zusammenfassung der Daten, sowohl nach Fehlerart als auch nach den verwendeten Anlagen:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *Beschreiben – Kategoriale Daten – Kreuztabellierung.*

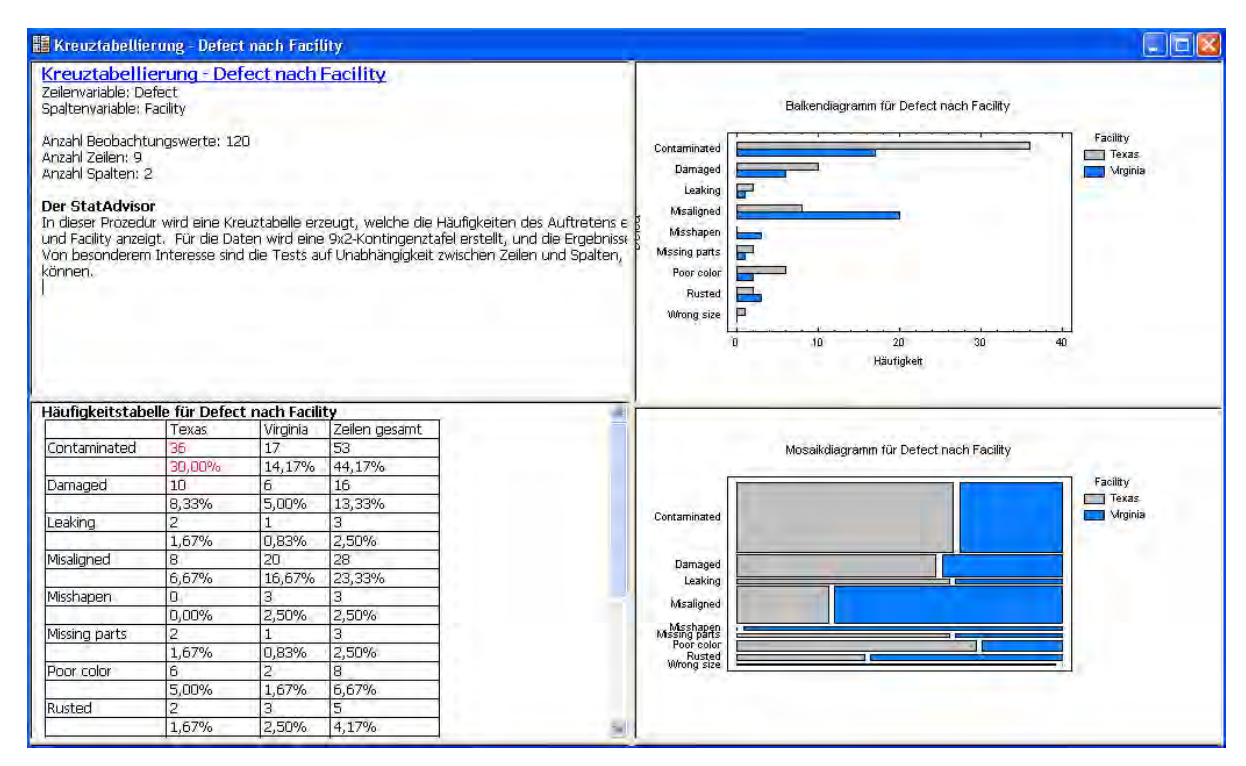
2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Attributive Daten – Mehrere Faktoren – Kreuztabellierung*.

Die Dateneingabe-Dialogbox benötigt nun zwei Variablen, eine zur Festlegung der Zeilen in der Kreuz- oder *Kontingenztabelle* und die andere zur Festlegung der Spalten:



Abbildung 14.6: Dateneingabe-Dialogbox für die Kreuztabellierung

Nach dem *Optionen*-Dialog und der Dialogbox *Tabellen und Grafiken* wird das folgende Analyse-Fenster erstellt:



	Texas	Virginia	Zeilen gesamt
Contaminated	36	17	53
	30,00%	14,17%	44,17%
Damaged	10	6	16
	8,33%	5,00%	13,33%
Leaking	2	1	3
	1,67%	0,83%	2,50%
Misaligned	8	20	28
	6,67%	16,67%	23,33%
Misshapen	0	3	3
	0,00%	2,50%	2,50%
Missing parts	2	1	3
	1,67%	0,83%	2,50%
Poor color	6	2	8
	5,00%	1,67%	6,67%
Rusted	2	3	5
	1,67%	2,50%	4,17%

Abbildung 14.7: Analyse-Fenster der Kreuztabellierung

Die Tabelle im unteren linken Ergebnisfenster enthält die gemeinsamen Häufigkeiten für *Defect* und *Facility*:

Häufigkeitstabelle für Defect nach Facility			
	Texas	Virginia	Zeilen gesamt
Contaminated	36	17	53
	30,00%	14,17%	44,17%
Damaged	10	6	16
	8,33%	5,00%	13,33%
Leaking	2	1	3
	1,67%	0,83%	2,50%
Misaligned	8	20	28
	6,67%	16,67%	23,33%
Misshapen	0	3	3
	0,00%	2,50%	2,50%
Missing parts	2	1	3
	1,67%	0,83%	2,50%
Poor color	6	2	8
	5,00%	1,67%	6,67%
Rusted	2	3	5
	1,67%	2,50%	4,17%
Wrong size	1	0	1
	0,83%	0,00%	0,83%
Spalten gesamt	67	53	120
	55,83%	44,17%	100,00%

Zellinhalt:
Beobachtete Häufigkeiten
Prozentwerte bezogen auf Tabelle

Abbildung 14.8: Kreuztabelle mit Prozentangaben, bezogen auf die gesamte Tabelle

Wie anfänglich gezeigt, enthält jede Zelle der Tabelle die Anzahl von Beobachtungen in der Datendatei, die zu dieser speziellen Zeile-Spalten-Kombination gehören. Sie enthält ebenfalls eine Prozentangabe zu dieser Zahl, bezogen auf die gesamte Tabelle. Zum Beispiel wurden in der Anlage *Texas* 36 Gegenstände mit dem Fehler *contaminated* hergestellt, das entspricht 30 % aller fehlerhaften Gegenstände in der Stichprobe.

Die *Ergebnisfenster-Optionen* ermöglichen die Auswahl anderer Zellinhalte:

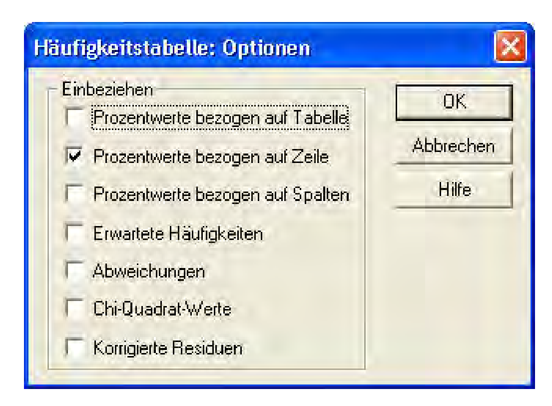


Abbildung 14.9: Ergebnisfenster-Optionen für die Kreuztabellierung

Eine interessante Variante für die aktuellen Daten ist die Anzeige der *Zeilenprozente* anstelle der *Tabellenprozente*:

Häufigkeitstabelle für Defect nach Facility

	Texas	Virginia	Zeilen gesamt
Contaminated	36	17	53
	67,92%	32,08%	44,17%
Damaged	10	6	16
	62,50%	37,50%	13,33%
Leaking	2	1	3
	66,67%	33,33%	2,50%
Misaligned	8	20	28
	28,57%	71,43%	23,33%
Misshapen	0	3	3
	0,00%	100,00%	2,50%
Missing parts	2	1	3
	66,67%	33,33%	2,50%
Poor color	6	2	8
	75,00%	25,00%	6,67%
Rusted	2	3	5
	40,00%	60,00%	4,17%
Wrong size	1	0	1
	100,00%	0,00%	0,83%
Spalten gesamt	67	53	120
	55,83%	44,17%	100,00%

Zellinhalt:
Beobachtete Häufigkeiten
Prozentwerte bezogen auf Zeilen

Abbildung 14.10: Kreuztabelle mit Prozentwerten, bezogen auf die Zeilen

Die Zeilenprozente zeigen den Prozentsatz an, den jede Zelle innerhalb ihrer Zeile verkörpert. Zum Beispiel wurden 67,92 % aller Teile mit dem Fehler *contaminated* in *Texas* produziert und 71,43 % aller Teile mit dem Fehler *misaligned* in *Virginia*. Das deutet darauf hin, dass manche Fehlerarten häufiger in der einen Anlage als in einer anderen vorkommen – eine Hypothese, die im nächsten Abschnitt formal getestet wird.

Verschiedene grafische Möglichkeiten sind ebenfalls sehr hilfreich. Zum Beispiel zeigt das Balkendiagramm die Daten für *Defect* nach *Facility*:

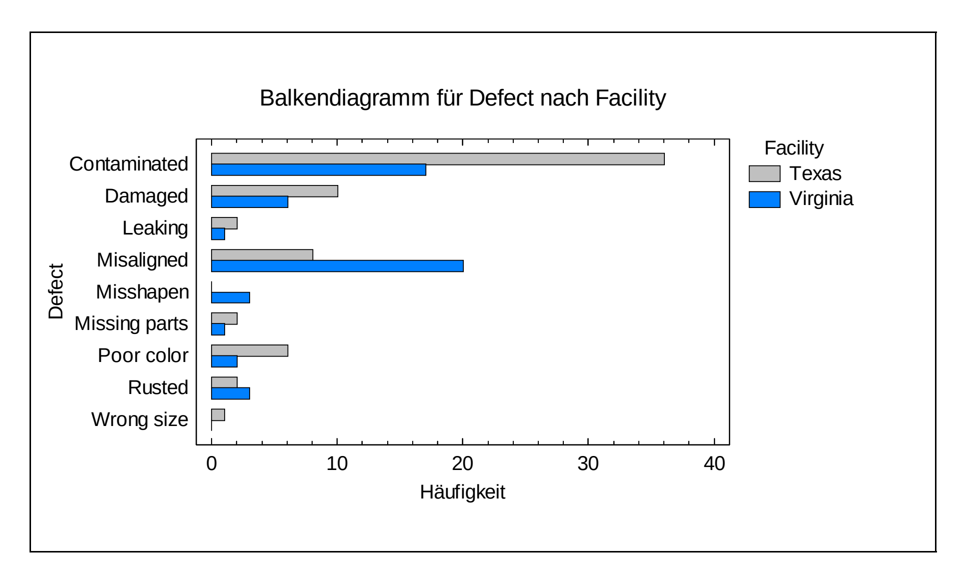


Abbildung 14.11: Gruppiertes Balkendiagramm

Die Unterschiede zwischen den beiden Anlagen sind ganz offensichtlich. Ein weiteres Diagramm, das so genannte *Mosaik-Diagramm*, ist ebenfalls sehr informativ:

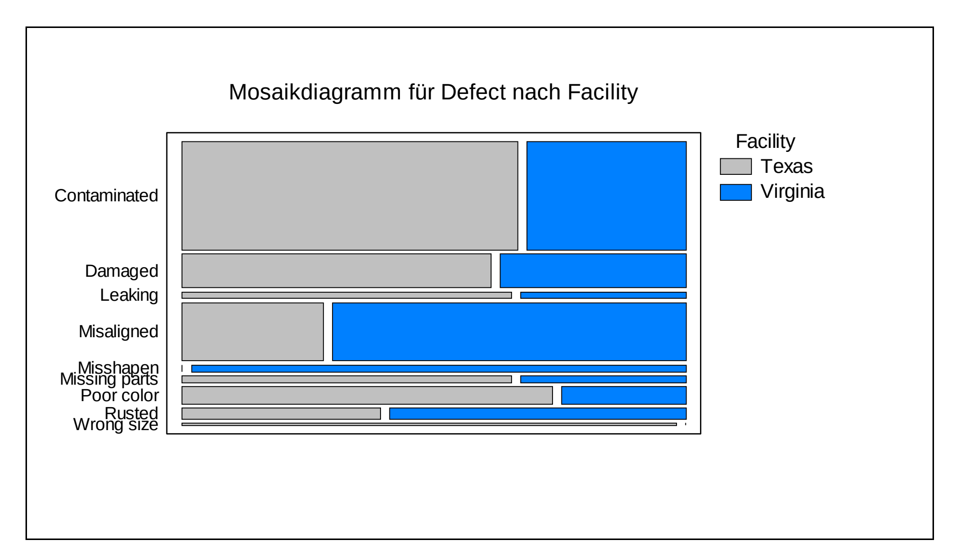


Abbildung 14.12: Mosaik-Diagramm

In diesem Diagramm ist die Höhe jedes Balkens proportional zur Gesamtzahl der Fehler je Fehlertyp. Die Breite der Balken ist proportional zur relativen Häufigkeit jedes Fehlertyps an den einzelnen Standorten. Folglich ist die Gesamtfläche jedes Rechtecks proportional zur Häufigkeit in der entsprechenden Zelle der Kreuztabelle.

Falls gewünscht, können die Zellenhäufigkeiten auch dreidimensional dargestellt werden, indem man in der Dialogbox *Tabellen und Grafiken* die Option *Sky-Diagramm* auswählt:

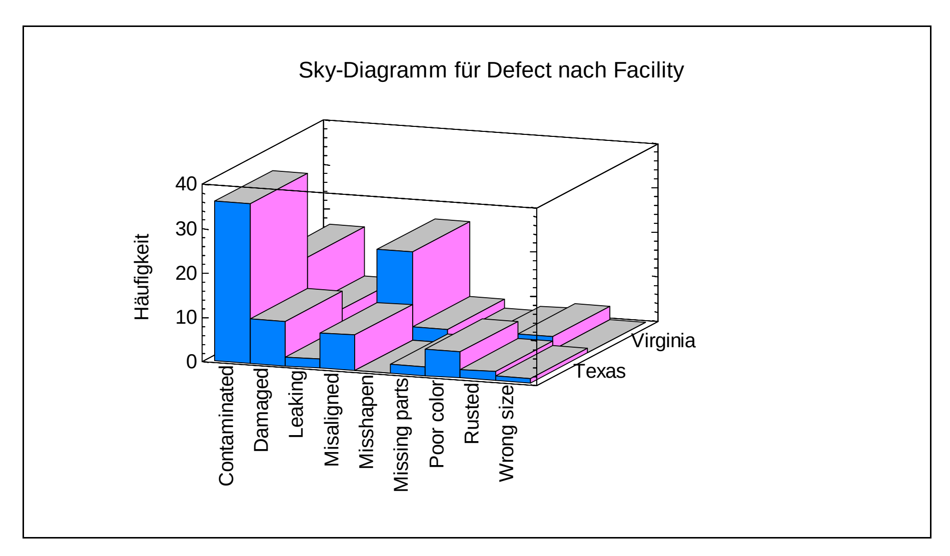


Abbildung 14.13: Dreidimensionales Sky-Diagramm

Im *Sky-Diagramm* repräsentiert die Höhe jedes Balkens die Zellhäufigkeit in der Kontingenztabelle.

14.4 Vergleichen von zwei oder mehr Stichproben

Um zu untersuchen, ob der scheinbare Unterschied zwischen den Anlagen in Texas und Virginia statistisch signifikant ist, wählt man die *Tests auf Unabhängigkeit* in der Dialogbox *Tabellen und Grafiken*. Bei einer Tabelle dieser Größe zeigt die Prozedur das Ergebnis eines Chi-Quadrat-Tests an:

Tests auf Unabhängigkeit

Test	*Statistik*	*FG*	*p-Wert*
Chi-Quadrat	18,438	8	0,0182

Warnung: einige Anzahlen pro Zelle < 5.

Abbildung 14.14: Chi-Quadrat-Test auf Unabhängigkeit

Der Chi-Quadrat-Test wird für folgende Hypothesen verwendet:

Null-Hypothese: Zeilen- und Spaltenklassifikation sind unabhängig.

Alternativ-Hypothese: Zeilen- und Spaltenklassifikation sind nicht unabhängig.

Unabhängigkeit würde hier bedeuten, dass die Fehlerart an einem Gegenstand nichts mit der Anlage zu tun hat, in der dieser Gegenstand hergestellt wurde.

Beim Chi-Quadrat-Test bedeutet ein kleiner *p*-Wert, dass die Zeilen- und Spaltenklassifikationen nicht unabhängig sind. In unserem Fall ist der *p*-Wert kleiner als 0,05, also ist die Verteilung der Fehlerarten in den Anlagen Texas und Virginia bei einer Irrtumswahrscheinlichkeit von 5 % signifikant verschieden.

Eine Warnung wird angezeigt, weil einige Anzahlen in den Zellen der Kreuztabelle kleiner als 5 sind. (Technisch taucht diese Warnung auf, wenn die erwarteten Anzahlen in irgendeiner Zelle kleiner als 5 sind, unter der Annahme, dass die Null-Hypothese zutrifft.) Bei kleinen Zellen-Anzahlen kann der *p*-Wert unzuverlässig sein. Eine mögliche Lösung dieses Problems ist die Zusammenfassung der seltenen Fehlerarten in eine einzige Klasse und die erneute Ausführung des Tests. In STATGRAPHICS Centurion XVI ist das sehr einfach auf folgendem Wege machbar:

1. Wechseln Sie in das Datenblatt und klicken Sie in den Spaltenkopf von *Defects*.

2. Drücken Sie die rechte Maustaste und wählen Sie im Kontextmenü *Daten rekodieren*.

3. Füllen Sie die Dialogbox *Daten rekodieren* aus, wie unten angezeigt, um alle weniger häufigen Fehlertypen in eine einzige Klasse mit der Beschriftung “Other” zusammenzufassen:



Abbildung 14.15: Rekodieren der weniger häufigen Fehlerarten

Die Eingaben in die Dialogbox *Daten rekodieren* veranlassen das Programm, nach Werten in der Spalte *Defects* zu suchen, welche in diese definierten Intervalle fallen. Jeder Wert, der innerhalb der Grenzen in einer bestimmten Zeile liegt, wird in den Wert rekodiert, der bei *Neuer Wert* eingetragen wurde.

Nach dem Ausführen der Rekodierung kehren Sie in das Analyse-Fenster der *Kreuztabellierung* zurück. Als Ergebnis der Änderungen im Datenblatt wurde die Analyse automatisch aktualisiert. Die neue Klasse *Other* besitzt jetzt eine ausreichend große Häufigkeit, wie im geänderten *Mosaik-Diagramm* zu sehen ist:

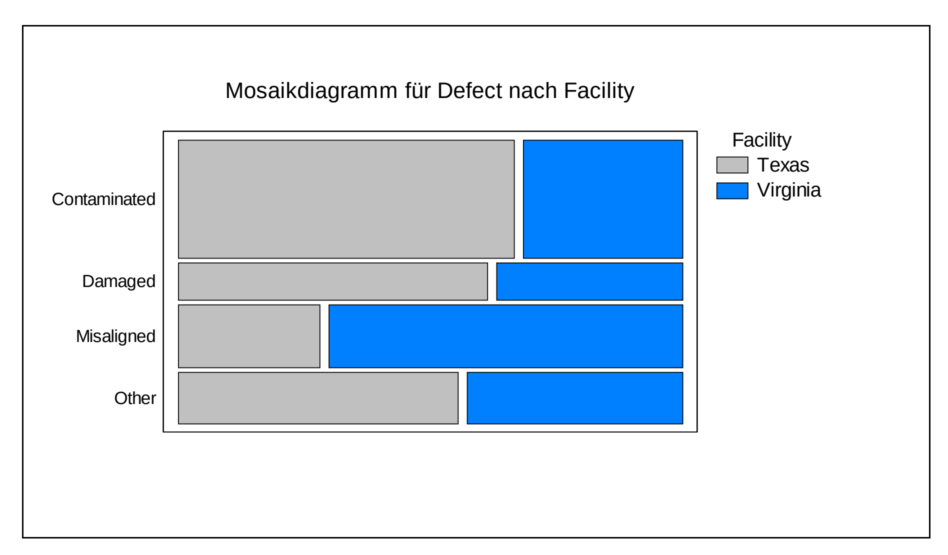


Abbildung 14.16: Mosaik-Diagramm für die rekodierten Daten

Nach dem Rekodieren zeigt der Chi-Quadrat-Test weiterhin einen statistisch signifikanten Unterschied zwischen den Anlagen Texas und Virginia an:

Tests auf Unabhängigkeit

Test	*Statistik*	*FG*	*p-Wert*
Chi-Quadrat	11,874	3	0,0078

Der StatAdvisor
Diese Tabelle zeigt das Ergebnis des Tests zur Überprüfung der Hypothese über die Unabhängigkeit der Zeilen- und Spaltenklassifikation. Da der p-Wert kleiner als 0,05 ist, können wir die Hypothese über die Unabhängigkeit zwischen Zeilen und Spalten bei einem Konfidenzniveau von 95,0% ablehnen. Folglich befindet sich der beobachtete Wert von Defect für einen speziellen Fall in Abhängigkeit von den Werten von Facility.

Abbildung 14.17: Chi-Quadrat-Test nach dem Rekodieren der Daten

Es sieht also tatsächlich so aus, dass der Fehlertyp im Zusammenhang mit der Anlage steht, in der produziert wurde.

Es sollte darauf hingewiesen werden, dass der obige Test die *Verteilung* der Fehlerarten für die beiden Anlagen vergleicht. Er vergleicht nicht die Anzahl oder Prozente an fehlerhaften Teilen an jedem Standort. Ein derartiger Vergleich erfordert einen anderen Test, wie im nächsten Abschnitt erklärt wird.

14.5 Kontingenztafeln

Um zu untersuchen, ob eine Anlage mehr fehlerhafte Teile produziert als eine andere, müssen Sie die gesamte Produktion jeder Anlage kennen. Angenommen, folgende Tabelle beschreibt die Produktion eines Monats:

Anlage	*Fehleranzahl*	*Anzahl produzierter Teile*
Texas	67	6237
Virginia	53	7343

Es sei θ_1 der Anteil fehlerhafter Teile, der in Texas produziert wurde, und θ_2 der Anteil fehlerhafter Teile, der in Virginia hergestellt wurde. Die geschätzten Anteile sind gegeben durch:

$$\hat{\theta}_1 = \frac{67}{6237} = 0{,}0107 \qquad \hat{\theta}_2 = \frac{53}{7343} = 0{,}0072$$

Auf Basis dieser Daten sieht es so aus, dass der Prozentsatz fehlerhafter Einheiten aus Texas größer sein könnte als der aus Virginia. Um festzustellen, ob dieser scheinbare Unterschied statistisch signifikant ist, erzeugen Sie folgendes Datenblatt:

<unbezeichnet>

	Attribut	Texas	Virginia
1	fehlerhaft	67	53
2	nicht fehlerhaft	6170	7290
3			
4			
5			

B C D E F G H

Abbildung 14.18: Datenblatt für den Vergleich zweier Prozentsätze

Die Zeilen enthalten die Anzahlen für fehlerhafte und nicht-fehlerhafte Teile. Danach wählen Sie die *Kontingenztabellen* im gleichen Menü wie die *Kreuztabellierung*. Geben Sie ein:

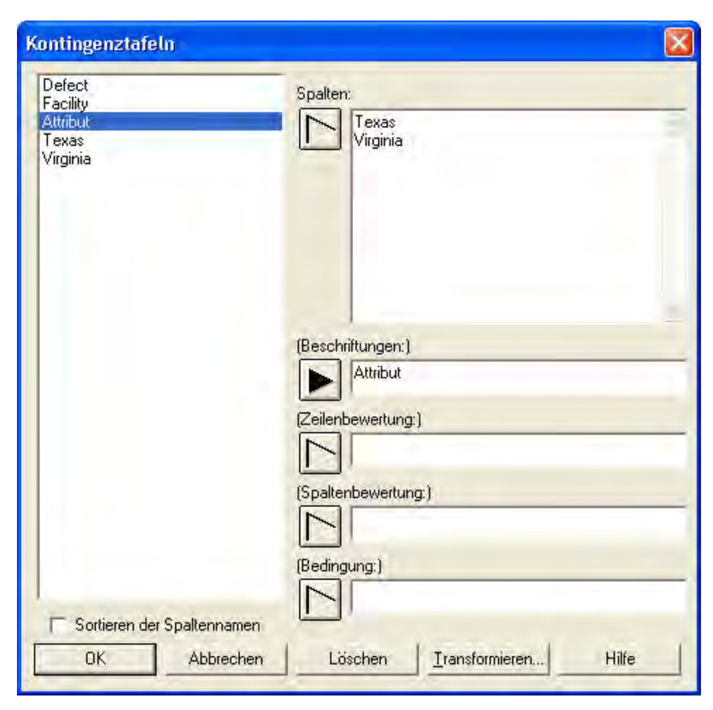


Abbildung 14.19: Dateneingabe-Dialogbox für Kontingenztafeln

Die Analyse zeigt den Chi-Quadrat-Test für eine Vierfeldertafel:

Tests auf Unabhängigkeit

Test	*Statistik*	*FG*	*p-Wert*
Chi-Quadrat	4,783	1	0,0287

Abbildung 14.20: Chi-Quadrat-Test für eine Vierfeldertafel

Erinnern Sie sich, dass der Chi-Quadrat-Test untersucht, ob die Zeilen- und Spaltenklassifikationen unabhängig sind. In unserem Fall würde eine Unabhängigkeit bedeuten, dass es nicht von der Produktionsanlage abhängt, ob ein Teil fehlerhaft ist oder nicht.
Da der *p*-Wert in obiger Tabelle kleiner als 0,05 ist, wird die Hypothese über die Unabhängigkeit bei einer Irrtumswahrscheinlichkeit von 5 % abgelehnt. Wir können daher schließen, dass die Anteile fehlerhafter Gegenstände für die beiden Anlagen signifikant verschieden sind.

Kapitel

15

Tutorial 6: Prozessfähigkeitsanalyse

Bestimmen der Defekte Pro Million (DPM) oder Prozent außerhalb der Spezifikationsgrenzen.

STATGRAPHICS Centurion XVI ist weit verbreitet unter den Anwendern, deren Aufgabe es ist, zu gewährleisten, dass die angebotenen Produkte und Services von höchster Qualität sind. Eine häufige Aufgabe bei diesen Jobs ist das Erfassen von Prozessdaten und deren Vergleich mit feststehenden Spezifikationsgrenzen. Das Ergebnis dieser Art von *Fähigkeitsanalysen* ist eine Schätzung darüber, wie ein Prozess in der Lage ist, solche Spezifikationen einzuhalten. Six Sigma, eine viel praktizierte Methode zur Erreichung von Weltniveau-Qualität, zielt auf eine Fehlerrate von 3,4 Fehler pro eine Million Möglichkeiten ab.

Betrachten Sie als Beispiel ein Produkt, für das eine Festigkeit (*strength*) zwischen 190 und 230 psi erforderlich ist. Angenommen, es wurden n = 100 Proben aus dem Fertigungsprozess entnommen und deren Festigkeit gemessen, wie in der folgenden Tabelle zu sehen ist:

213,5	203,3	191,3	197,1	205,7	215,6	193,7	201,7	201,5	207,1
207,0	200,4	197,2	202,4	205,2	211,0	214,5	201,5	200,9	206,8
205,8	200,3	196,1	205,9	195,1	203,9	192,9	199,0	195,5	203,1
197,4	194,8	201,0	202,5	199,0	200,7	197,6	198,5	205,3	197,1
202,8	201,6	197,4	200,9	203,3	209,4	201,4	199,5	207,8	204,9
205,5	203,0	208,1	200,2	218,2	202,0	209,3	201,2	200,4	201,0
195,7	229,5	199,9	208,1	210,3	202,0	202,6	213,6	198,0	197,8
196,7	216,0	211,6	208,7	199,4	200,8	201,1	195,3	206,8	211,3
201,5	200,0	211,8	195,6	201,9	199,0	200,3	197,8	200,8	194,8
199,5	195,5	201,0	206,0	215,3	202,6	199,9	200,6	197,6	207,4

In diesem Kapitel wird beschrieben, wie man eine typische Fähigkeitsanalyse für diese Art von messbaren Daten durchführt.

15.1 Grafische Darstellung der Daten

Der erste Schritt bei der Untersuchung einer jeden neuen Datenmenge ist die grafische Darstellung. Für eine Datenmenge wie die oben gezeigte bietet die *Analyse einer Variablen*, die in Kapitel 10 beschrieben wurde, eine Reihe nützlicher Werkzeuge. Um die Daten zu analysieren:

1. Öffnen Sie die Datei *items.sgd.*

2. Führen Sie die Prozedur *Analyse einer Variablen* für die Spalte *Strength* durch.

Das anfängliche Analyse-Fenster sieht folgendermaßen aus:

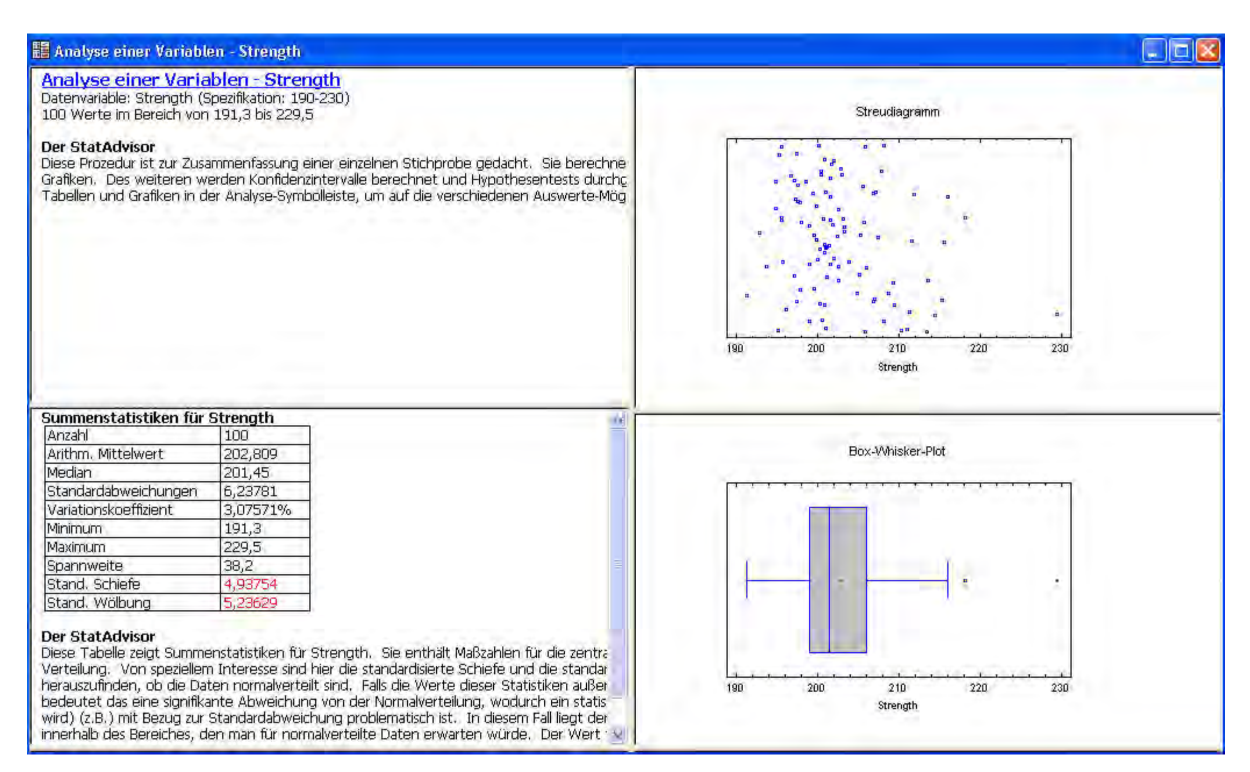


Abbildung 15.1: Analyse-Fenster für die Analyse einer Variablen

Mehrere interessante Fakten sind sofort offensichtlich:

1. Die Daten liegen in einem Bereich von 191,3 bis 229,5 und damit, wenn auch knapp, innerhalb der Spezifikationsgrenzen.

2. Das Box-Whisker-Plot zeigt einen weit außerhalb liegenden Punkt an (kleines Quadrat mit einem roten Pluszeichen darin). Solche Punkte werden häufig als Ausreißer betrachtet, wenn der Rest der Daten normalverteilt aussieht. In unserem Fall jedoch ist die Gestalt der Box nicht sehr symmetrisch, selbst wenn der scheinbare Ausreißer nicht berücksichtigt wird. Der obere Ausläufer (Whisker) ist länger als der untere, und die Box erstreckt sich oberhalb des Medians (vertikale Linie in der Box) weiter als unterhalb.

3. Wenn Sie das Ergebnisfenster *Summenstatistiken* anschauen, werden Sie feststellen, dass die standardisierte Schiefe 4,94 ist. Falls die Daten normalverteilt sind, sollte dieser Wert zwischen -2 und +2 liegen. Selbst die Entfernung des größten Wertes reduziert die standardisierte Schiefe nur auf 2,81.

Durch Anklicken der Schaltfläche *Tabellen und Grafiken* auf der Analyse-Symbolleiste und Auswahl von *Häufigkeitsdiagramm* kann man außerdem ein *Histogramm* anzeigen lassen:

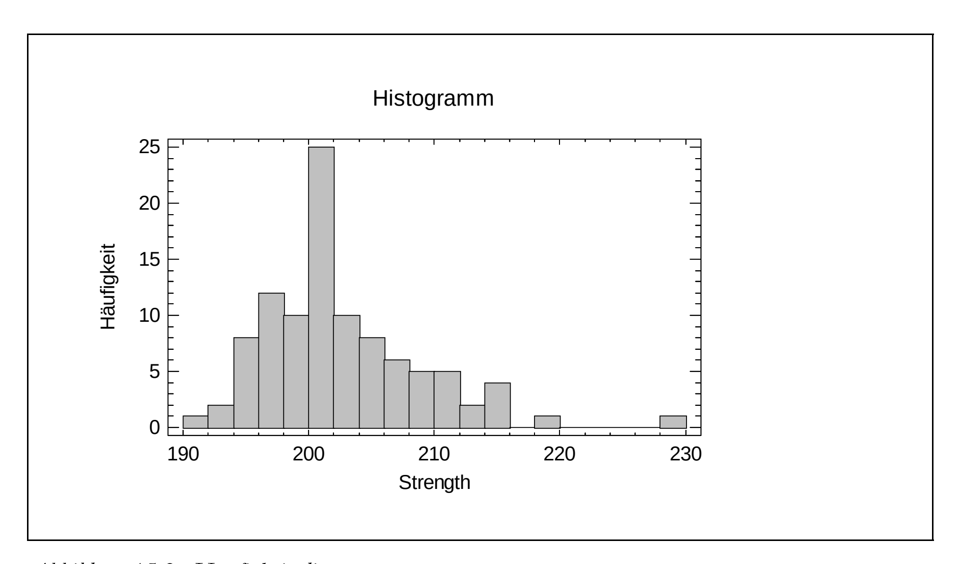


Abbildung 15.2: Häufigkeitsdiagramm

Die Daten scheinen ganz klar rechtsschief zu sein, indem sie sich vom Gipfel ausgehend weiter nach rechts ausdehnen als nach links.

Nicht-normalverteilte Daten wie die gezeigten sind eine Alltäglichkeit. Leider kommt es im Umgang mit solchen Daten häufig vor, dass die Nicht-Normalverteiltheit schlicht ignoriert und

die Indizes wie C_{pk} mit Formeln berechnet werden, die eigentlich für normalverteilte Daten gedacht sind. Wie in diesem Tutorial zu sehen ist, führt das Ignorieren der Nicht-Normalverteiltheit zu inkorrekten Ergebnissen, die häufig den Prozentwert von außerhalb der Spezifikation liegenden Produkten über- oder unterschätzen.

15.2 Prozessfähigkeitsanalyse

STATGRAPHICS Centurion XVI enthält Prozeduren zur Durchführung einer Fähigkeitsanalyse für Daten, die entweder einen Wert pro Zeitpunkt (Einzelwerte) oder Werte in Untergruppen sammeln (z. B. jede Stunde 5 Beobachtungen). Angenommen, es handelt sich um Einzelwerte, dann kann die Prozessfähigkeitsanalyse folgendermaßen durchgeführt werden:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *SPC – Fähigkeitsanalyse – Messbare Merkmale – Einzelwerte.*

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Analysieren – Messbare Merkmale– Fähigkeitsanalyse – Einzelwerte.*

Die Dateneingabe-Dialogbox verlangt den Namen einer einzigen Spalte, welche alle Daten enthält. Als Beispiel kann die Spalte *Strength* aus der Datei *items.sgd* verwendet werden:

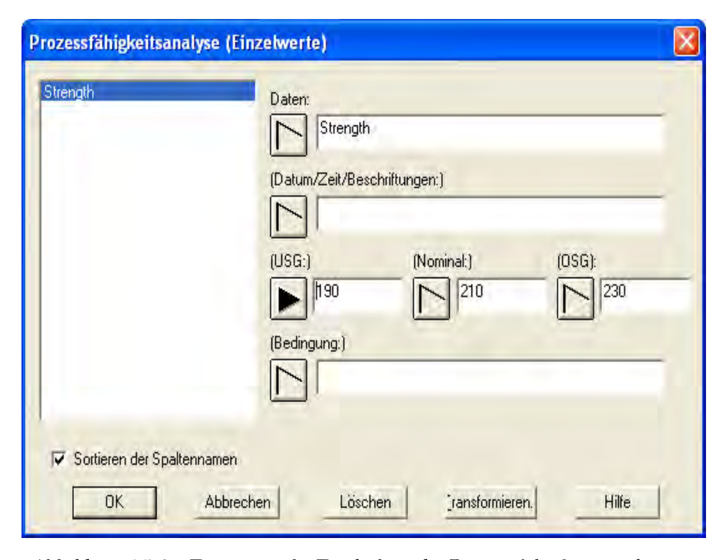


Abbildung 15.3: Dateneingabe-Dialogbox der Prozessfähigkeitsanalyse

Obere und untere Spezifikation sind ebenfalls anzugeben, ebenso ein Nominal- oder Zielwert.

Nach dem Anklicken von *OK* wird die Dialogbox *Optionen* und anschließend die Dialogbox *Tabellen und Grafiken* geöffnet. Für unser Beispiel können Sie die voreingestellten Standardeinstellungen verwenden.

Das anfängliche Analyse-Fenster zeigt eine Datenzusammenfassung an sowie die Tabelle der Fähigkeitsindizes und das Fähigkeitsdiagramm:

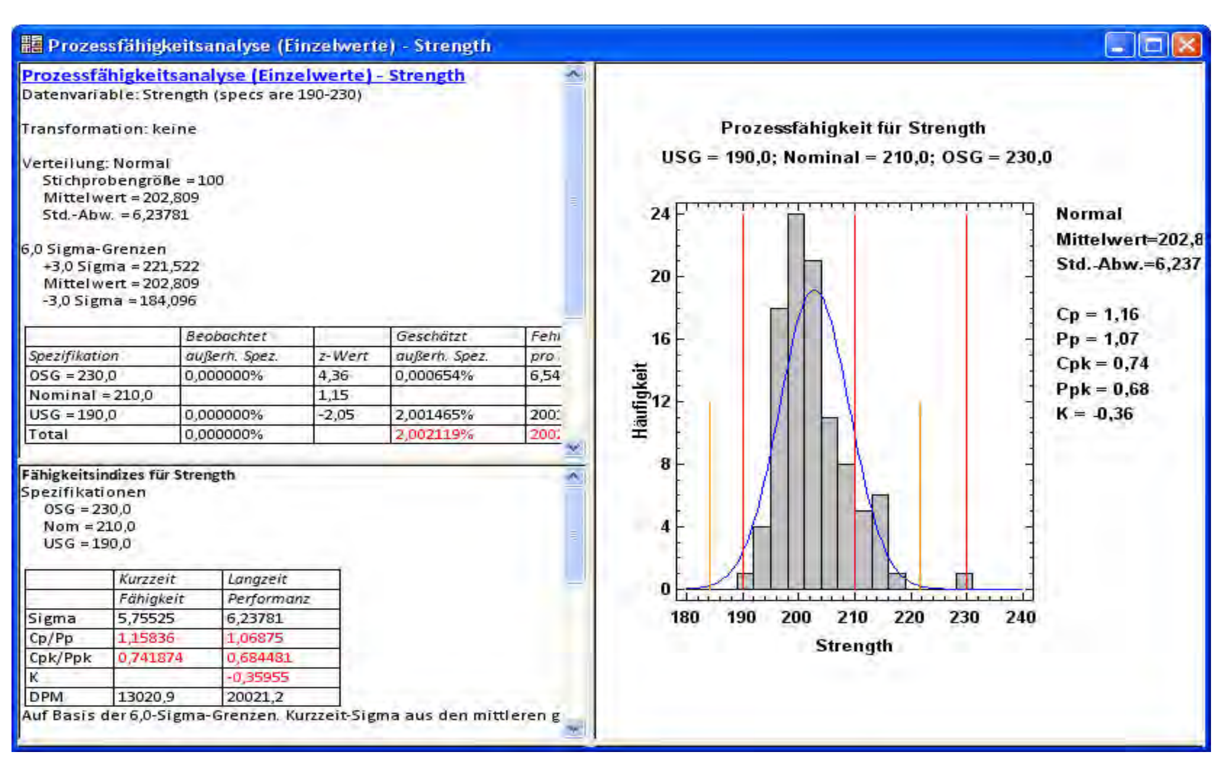


Abbildung 15.4: Analyse-Fenster der Prozessfähigkeitsanalyse

Wenn eine Prozessfähigkeitsanalyse das erste Mal abläuft, wird zunächst eine Normalverteilung an die Daten angepasst. Das *Fähigkeitsdiagramm* zeigt ein Histogramm der Daten, zusammen mit der besten angepassten Normalverteilung:

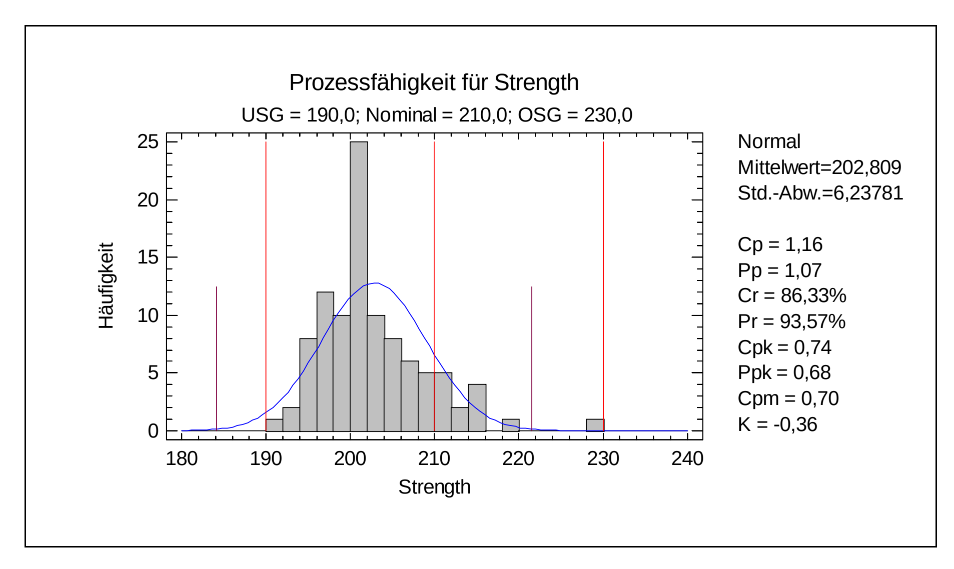


Abbildung 15.5: Fähigkeitsdiagramm mit Normalverteilung

Die langen vertikalen Linien zeigen die Lage der Spezifikationsgrenzen und des Nominalwertes an. Die kürzeren vertikalen Linien sind an den Positionen des Stichproben-Mittelwerts +/- 3 Standardabweichungen lokalisiert. Zum obigen Diagramm ist insbesondere Folgendes bemerkenswert:

1. Die angepasste Normalverteilung ist für die Daten nicht geeignet. Obwohl die Normalverteilung mit Mittelwert und Standardabweichung, die aus den Daten stammen, angepasst wurde, so sieht man doch, dass aufgrund der Schiefe in den Daten keine gute Übereinstimmung zwischen den Histogramm-Säulen und der Glockenkurve besteht.

2. Der Stichproben-Mittelwert liegt bei 202,8, was bedeutend weniger als der Nominalwert von 210 ist.

3. Obwohl keine Beobachtung kleiner ist als die untere Spezifikationsgrenze, liegt ein gutes Stück vom unteren Schwanz der Normalverteilung unterhalb dieser Grenze.

4. Die Linien bei +/- 3 Sigma sind zwar eng genug zusammen, so dass sie zwischen die Spezifikationsgrenzen passen würden, sie sind jedoch nach links verschoben.

Die *Analyse-Zusammenfassung* im oberen linken Ergebnisfenster quantifiziert die Anpassung:

Prozessfähigkeitsanalyse (Einzelwerte) - Strength
Datenvariable: Strength (Spezifikation: 190-230)

Transformation: keine

Verteilung: Normal
Stichprobengröße = 100
Mittelwert = 202,809
Std.-Abw. = 6,23781

6,0 Sigma-Grenzen
+3,0 Sigma = 221,522
Mittelwert = 202,809
-3,0 Sigma = 184,096

	Beobachtet		*Geschätzt*	*Fehler*
Spezifikation	*außerh. Spez.*	*z-Wert*	*außerh. Spez.*	*pro Million*
OSG = 230,0	0,000000%	4,36	0,000654%	6,54
Nominal = 210,0		1,15		
USG = 190,0	0,000000%	-2,05	2,001465%	20014,65
Total	0,000000%		2,002119%	20021,19

Abbildung 15.6: Zusammenfassung der Analyse zur Prozessfähigkeit

Von erstrangigem Interesse ist die untere Tabelle, in der die Prozente von außerhalb der Spezifikationen liegenden Produkten geschätzt werden. Auf Basis der aktuell angepassten Normalverteilung würden geschätzte 2 % der Produkte außerhalb der Spezifikationsgrenzen liegen, das wären 20 021 Defekte pro Million (DPM).

15.3 Umgang mit nicht-normalverteilten Daten

Die geschätzten DPM, die oben berechnet wurden, hängen in starken Maße von der Voraussetzung ab, dass die Daten normalverteilt sind. Ein formaler Test über die Gültigkeit dieser Voraussetzung kann nach Auswahl von *Tests auf Normalverteilung* in der Dialogbox *Tabellen und Grafiken* ausgeführt werden:

Tests auf Normalverteilung für Strength

Test	*Statistik*	*p-Wert*
Shapiro-Wilks-W	0,931784	0,0000321356

Abbildung 15.7: Test auf Normalverteilung

In Abhängigkeit von Ihren System-Voreinstellungen werden ein oder mehrere Tests auf Normalverteilung angezeigt. Jeder dieser Tests basiert auf den folgenden Hypothesen:

Null-Hypothese: Daten sind normalverteilt.

Alternativ-Hypothese: Daten sind nicht normalverteilt.

Ein *p*-Wert kleiner als 0,05 führt zur Ablehnung der Hypothese über die Normalverteilung bei einer Irrtumswahrscheinlichkeit von 5 %.

In der obigen Tabelle ist eine deutliche Ablehnung der Hypothese, dass die Daten normalverteilt sind, durch den Shapiro-Wilks-Test zu erkennen. Folglich ist jeder auf Basis der Normalverteilung geschätzte DPM-Wert oder Fähigkeitsindex nicht korrekt.

Sofern Daten nicht normalverteilt sind, kann man auf einem der beiden weiteren möglichen Wege fortfahren:

1. Auswahl einer anderen Verteilung, auf deren Basis dann die Analyse durchgeführt wird.

2. Transformation der Daten, so dass die transformierten Daten möglicherweise normalverteilt sind.

Zur Unterstützung bei der Auswahl einer anderen Verteilung bietet STATGRAPHICS Centurion XVI die Option *Vergleich der alternativen Verteilungen* in der Dialogbox *Tabellen und Grafiken* an. Hier werden eine Reihe anderer Verteilungen angepasst und nach der Anpassungsgüte sortiert aufgelistet. Unter Verwendung der Standardauswahl von Verteilungen wird folgendes Ergebnis erzeugt:

Vergleich der alternativen Verteilungen

Verteilung	*Geschätzte Parameter*	*Log Likelihood*	*KS D*	*A^2*
Extremwerte für Maxima	2	-314,65	0,0675422	0,372613
Laplace	2	-320,055	0,0920985	1,68399
Loglogistische	2	-320,271	0,0913779	1,15081
Logistische	2	-321,236	0,0941708	1,27599
Lognormal	2	-322,763	0,13213	1,66564
Gamma	2	-323,306	0,134136	1,73401
Normal	2	-324,457	0,138628	1,90094
Weibull	2	-348,002	0,177886	5,67166
Extremwerte für Minima	2	-351,782	0,189989	6,28546
Exponential	1	-631,226	0,61064	43,3327
Pareto	1	-798,174	0,628084	45,3859

Abbildung 15.8: Angepasste Verteilungen sortiert nach der Anpassungsgüte

Die Verteilungen sind nach dem Wert der Kolmogorov-Smirnov-Statistik sortiert, die den maximalen Abstand zwischen der kumulativen Verteilung der Daten und der theoretischen Verteilung misst. In unserem Fall ist die Extremwertverteilung für Maxima die am besten angepasste Verteilung.

Man kann unter Verwendung der *Analyse-Optionen* zur Extremwertverteilung für Maxima wechseln:

Abbildung 15.9: Analyse-Optionen der Prozessfähigkeitsanalyse

Das Anpassungsergebnis ist folgendes:

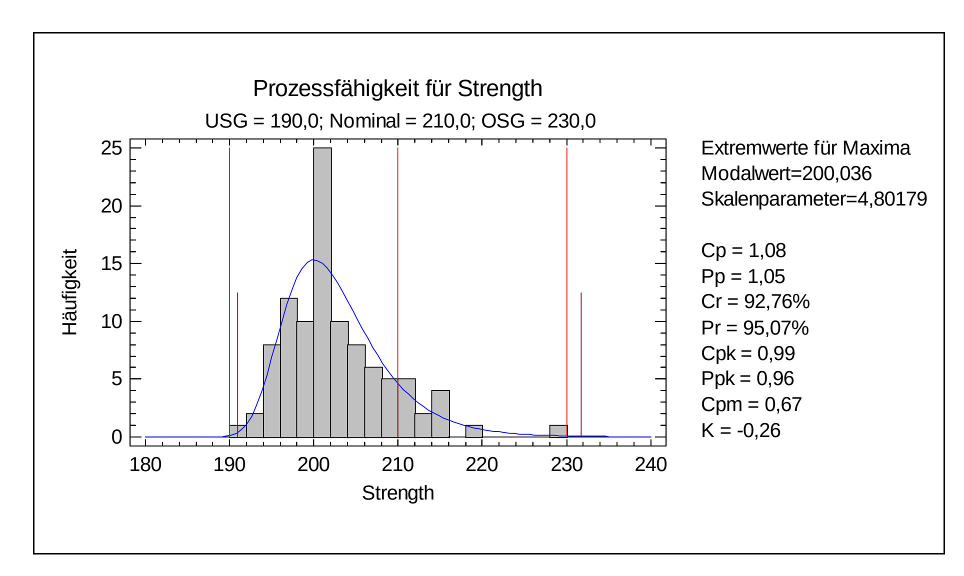


Abbildung 15.10: Angepasste Extremwertverteilung für Maxima

Beachten Sie, dass die Verteilung rechtsschief ist und mit den Beobachtungsdaten viel besser als die Normalverteilung übereinstimmt. Die kurzen vertikalen Linien sind bei den "äquivalenten" drei Sigma-Grenzen angeordnet, d. h. innerhalb dieser Grenzen liegen 99,73 % der angepassten Verteilung, das entspricht bei der Normalverteilung innerhalb der Grenzen Mittelwert plus/minus 3 Sigma. Diese Grenzen sind wegen der positiven Schiefe nicht symmetrisch um die Spitze der Verteilung angeordnet.

Die *Analyse-Zusammenfassung* zeigt eine drastische Differenz bezüglich der geschätzten Prozente an außerhalb der Spezifikation liegenden Produkten, wenn man das mit der vorhin angepassten Normalverteilung vergleicht:

Prozessfähigkeitsanalyse (Einzelwerte) - Strength
Datenvariable: Strength (Spezifikation: 190-230)

Transformation: keine

Verteilung: Extremwerte für Maxima
Stichprobengröße = 100
Modalwert = 200,036
Skalenparameter = 4,80179
(Mittelwert = 202,808)
(Sigma = 6,15853)

Äquivalent 6,0 Sigma-Grenzen
99,865 Perzentil = 231,761
Median = 201,796
0,134996 percentile = 190,969

	Beobachtet		*Geschätzt*	*Fehler*
Spezifikation	*außerh. Spez.*	*z-Wert*	*außerh. Spez.*	*pro Million*
OSG = 230,0	0,000000%	2,89	0,194758%	1947,58
Nominal = 210,0		1,19		
USG = 190,0	0,000000%	-3,42	0,030805%	308,05
Total	0,000000%		0,225563%	2255,63

Abbildung 15.11: Analyse-Zusammenfassung nach Anpassung der Extremwertverteilung für Maxima

Die geschätzten Prozent außerhalb der Spezifikation sind jetzt nur 0,23 oder 2.256 DPM, d. h. ein Zehntel von dem, was unter Verwendung der Normalverteilung berechnet wurde. In diesem Fall lässt die inkorrekte Annahme der Normalverteilung den Prozess deutlich schlechter erscheinen als er tatsächlich ist.

> ANMERKUNG: In Abhängigkeit von den Spezifikationsgrenzen und der wahren Verteilung kann die inkorrekte Annahme der Normalverteilung den Prozess signifikant schlechter oder signifikant besser erscheinen lassen als bei Verwendung der richtigen Verteilung.

Die Alternative zur Verwendung einer anderen Verteilung ist eine Datentransformation. Die Dialogbox *Analyse-Optionen* bietet verschiedene Möglichkeiten der *Datentransformation*:

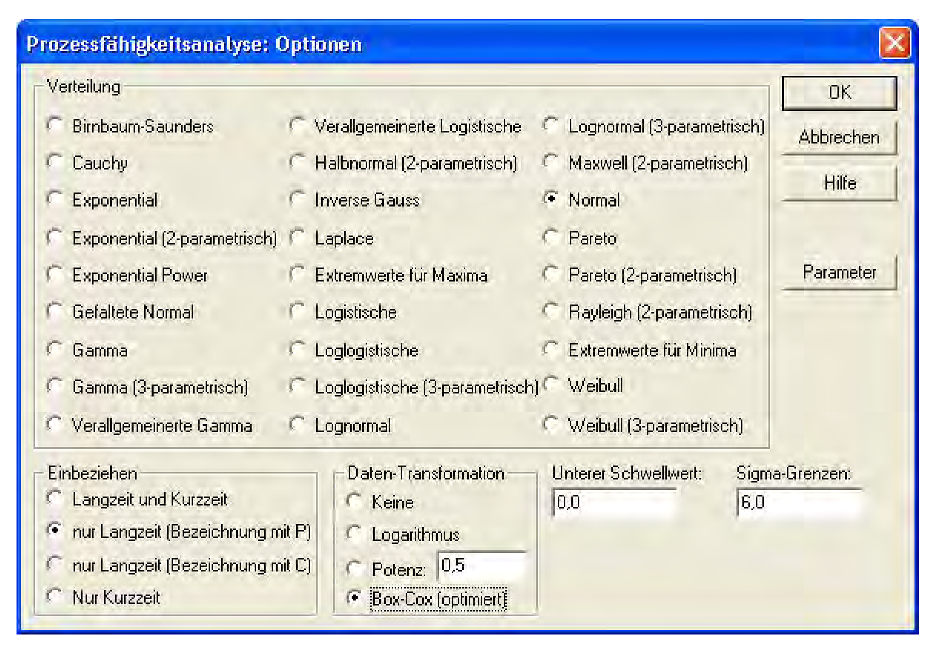


Abbildung 15.12: Analyse-Optionen zur Auswahl von Datentransformationen

Zur Auswahl stehen der natürliche Logarithmus, das Berechnen der Potenz mit einem festgelegten Exponenten oder eine Box-Cox-Transformation. Die letzte Variante zieht eine Reihe von Transformationen der Form Y^p in Erwägung, wobei der optimale Wert für p mittels Box-Cox-Methoden ermittelt wird.

Nachdem eine Transformation gewählt wurde, wird eine Normalverteilung an die transformierten Daten angepasst. Das folgende Diagramm zeigt die Ergebnisse nach Anwendung der Box-Cox-Transformation:

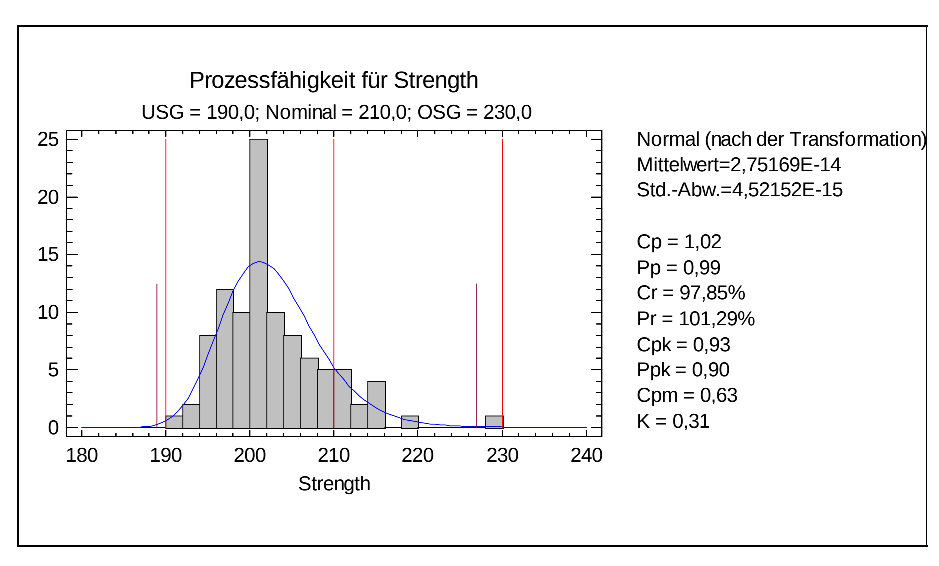


Abbildung 15.13: Fähigkeitsdiagramm nach der Box-Cox-Transformation

Für das Diagramm wurde eine inverse Transformation angewendet, um die Anpassung in der originalen Metrik zu zeigen. Die Transformation hatte einen ähnlichen Effekt auf die Gestalt der Verteilung, jedoch nicht so stark, wie die Annahme der Extremwertverteilung für Maxima. Der geschätzte DPM-Wert ist 4.353, das ist etwa zweimal so groß wie bei Verwendung der Extremwertverteilung für Maxima, jedoch deutlich kleiner als unter der Annahme der Normalverteilung.

> ANMERKUNG: Mittelwert und Standardabweichung, die in dem Diagramm angezeigt werden, gehören zu den transformierten Daten und sind im Allgemeinen nicht besonders nützlich. STATGRAPHICS Centurion XVI konvertiert deshalb automatisch alles zurück in die Originalgrößen.

Um die beiden Varianten zu vergleichen, kann das *Wahrscheinlichkeitsdiagramm* in der Dialogbox *Tabellen und Grafiken* für jede Variante gewählt und nebeneinander in die StatGallery eingefügt werden:

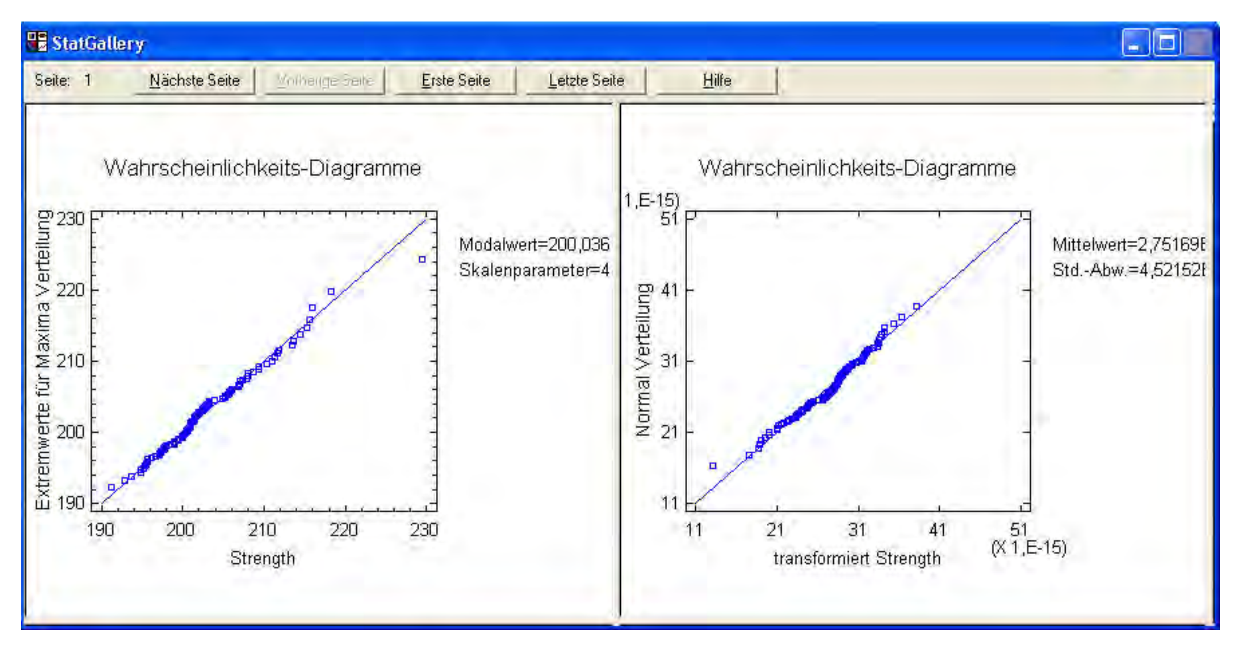


Abbildung 15.14: Wahrscheinlichkeitsdiagramme in der StatGallery

Wenn die angenommene Verteilung korrekt ist, sollten die Punkte bei Darstellung in diesem Diagramm entlang der Diagonalen liegen. Beide Varianten scheinen die Nicht-Normalverteiltheit gut bewältigt zu haben, was es schwer macht, zwischen ihnen zu wählen. Welche Methode auch immer verwendet wird, es ist wichtig, einen Standard dafür einzurichten, welche spezielle Variable (z. B. *Strength*) wie behandelt wird und diesen dann immer für die Analyse derartiger Daten anzuwenden. Es wäre falsch, diese Art von explorativer Datenanalyse, die in diesem Kapitel beschrieben wird, jedes Mal durchzuführen, wenn eine Menge ähnlicher Daten gesammelt wurde. Stattdessen sollte diese Analyseart einmal gemacht werden, um zu untersuchen, wie die spezielle Variable zu behandeln ist, und dann sollte die gewählte Variante auf die Variable angewendet werden, wann immer sie analysiert wird.

15.4 Fähigkeitsindizes

Das Wesentliche der Fähigkeitsanalyse besteht darin, den Prozentsatz von Produkten zu schätzen, die außerhalb der Spezifikationsgrenzen liegen würden (oder äquivalent DPM, Defekte auf 1 Million). Zur Zusammenfassung der Prozessfähigkeit haben Praktiker eine Reihe verschiedener Fähigkeitsindizes abgeleitet. Der am häufigsten berechnete Index ist C_{pk}, definiert als:

$$C_{pk} = \min\left(\frac{\hat{\mu} - USG}{3\hat{\sigma}}, \frac{OSG - \hat{\mu}}{3\hat{\sigma}}\right)$$

Einfach ausgedrückt ist C_{pk} der Abstand vom geschätzten Prozessmittelwert zur nächstgelegenen Spezifikationsgrenze, dividiert durch dreimal die geschätzte Prozess-Standardabweichung.

Die Prozedur *Prozessfähigkeitsanalyse* in STATGRAPHICS Centurion XVI zeigt die Indizes sowohl auf dem *Fähigkeitsdiagramm* als auch in der Tabelle der *Fähigkeitsindizes* an. Zunächst unter Annahme der Normalverteilung werden sowohl Langzeit- als auch Kurzzeit-Fähigkeiten berechnet.

Fähigkeitsindizes für Strength
Spezifikationen
OSG = 230,0
Nom = 210,0
USG = 190,0

	Kurzzeit	*Langzeit*
	Fähigkeit	*Performanz*
Sigma	5,75525	6,23781
Cp/Pp	1,15836	1,06875
CR/PR	86,3287	93,5672
Cpk/Ppk	0,741874	0,684481
Cpk/Ppk (oberes)	1,57485	1,45302
Cpk/Ppk (unteres)	0,741874	0,684481
Cpm		0,698308
K		-0,35955
% außerhalb der Spez.	1,30209	2,00212
DPM	13020,9	20021,2
Sigma-Qualitätsstufe	3,72559	3,55332

Auf Basis der 6,0-Sigma-Grenzen. Kurzzeit-Sigma aus den mittleren gleitenden Spannweiten geschätzt. Die Sigma-Qualitätsstufe beinhaltet eine 1,5-Sigma-Veränderung im Mittelwert.

95,0% Konfidenzintervalle

Index	*Untere Grenze*	*Obere Grenze*
Cp	0,997149	1,31931
Pp	0,920008	1,21725
Cpk	0,619618	0,864129
Ppk	0,568904	0,800059
Cpm	0,61885	0,777645

Abbildung 15.15: Tabelle der Fähigkeitsindizes

Die Kurzzeit-Fähigkeitsindizes, die unter Verwendung einer Schätzung von Sigma aus zeitlich nahe beieinander liegenden Beobachtungen berechnet wird, beschreibt die „Fähigkeit" des Prozesses bei konstant gehaltenem Mittelwert. Die Langzeit-Indizes werden mit Hilfe einer Sigma-Schätzung von der gesamten Variabilität der Beobachtungen innerhalb des ganzen Stichprobenzeitraumes berechnet. Sie beschreiben, wie der Prozess tatsächlich ausgeführt wird: die Performanz des Prozesses. Ein Prozess, der außer Kontrolle ist, in dem sich der Mittelwert während des Datensammelns signifikant verändert hat, kann eine wesentlich schlechtere Performanz zeigen als ein Prozess, der sich unter Kontrolle befindet. Standardmäßig bezeichnet STATGRAPHICS Centurion XVI die Fähigkeitsindizes mittels des Buchstabens „C" und die Performanzindizes mittels des Buchstabens „P".

Im Register *Fähigkeit* der Dialogbox *Voreinstellungen*, welche unter *Bearbeiten* im STATGRAPHICS Centurion XVI-Hauptmenü zu finden ist, kann man die Indizes festlegen, die standardmäßig berechnet werden sollen, ebenso wie andere wichtige Optionen:

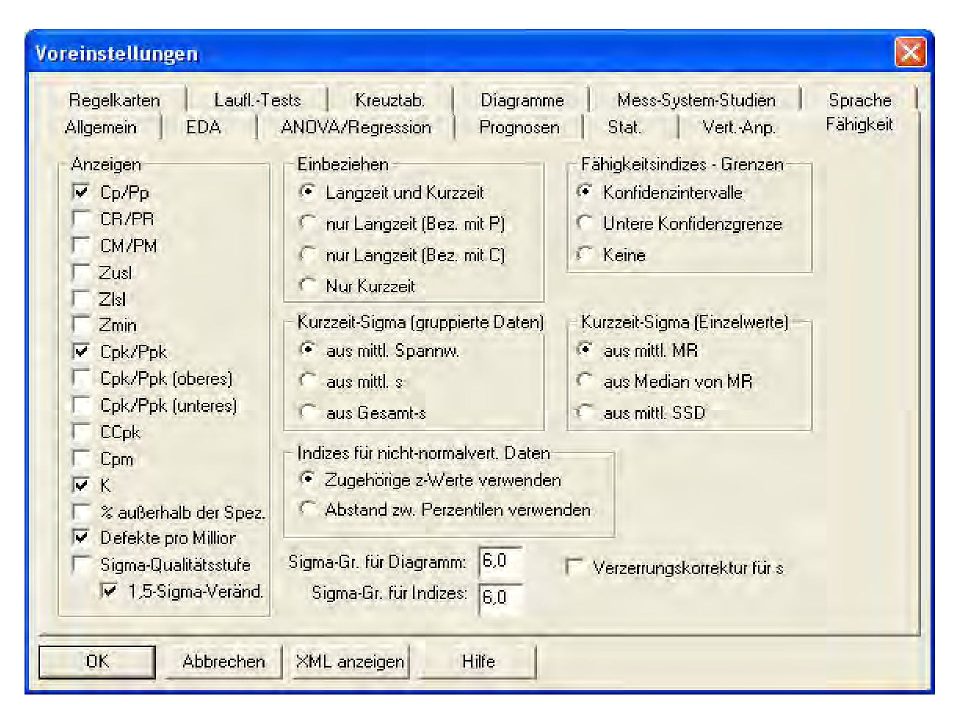


Abbildung 15.16: System-Voreinstellungen für Fähigkeitsindizes

Die linke Seite in der Dialogbox listet die Indizes auf, die man berechnen kann. Zusätzlich zu C_{pk}, sind darunter u.a. folgende Indizes:

1. C_p – ein zweiseitiger Fähigkeitsindex, berechnet aus

$$C_p = \frac{OSG - USG}{6\hat{\sigma}}$$

Dieser Index misst den Abstand zwischen den Spezifikationsgrenzen relativ zum Bereich, der durch sechs Standardabweichungen überdeckt wird. C_p ist immer größer oder gleich C_{pk}. Eine wesentliche Differenz zwischen den beiden Indizes zeigt an, dass der Prozess nicht gut zentriert ist.

2. *K* – ein Maß dafür, wie weit der Prozess vom Zentrum entfernt ist. *K* wird berechnet aus

$$K = \frac{\hat{\mu} - NOM}{(OSG - USG)/2}$$

wobei *NOM* der Nominal- oder Zielwert ist. Ein Wert von *K* nahe 0 ist ein Zeichen dafür, dass der Prozess gut zentriert ist.

3. *Sigma-Qualitätsniveau* – ein Index, welcher in der Six-Sigma-Methode benutzt wird, um das mit dem Prozess verknüpfte Qualitätsniveau anzuzeigen. Ein *Sigma-Qualitätsniveau* von 6 entspricht normalerweise einem DPM-Wert von 3,4 Defekten pro Million.

Die Dialogbox *Voreinstellungen* beeinflusst auch, welche Indizes auf dem *Fähigkeitsdiagramm* angezeigt werden und wie sie beschriftet sind. Eine detaillierte Diskussion der verschiedenen Indizes kann man im PDF-Dokument mit dem Titel *Capability Analysis – Variable Data (Fähigkeitsanalyse – Messbare Daten)* finden.

Zusätzlich zu den Fähigkeitsindizes enthält die Tabelle in Abbildung 15.14 Konfidenzintervalle, welche die Genauigkeit der Schätzung für diese Indizes anzeigen. Zum Beispiel zeigt die obige Tabelle einen C_{pk}–Wert von 0,74. Das 95%-Konfidenzintervall reicht von 0,62 bis 0,86. Das bedeutet, der wahre Wert von C_{pk} für den Prozess, aus dem die Daten stammen, liegt irgendwo zwischen 0,62 und 0,86.

Wenn die Daten nicht normalverteilt sind, müssen die Fähigkeitsindizes modifiziert werden. Nach der Standardeinstellung in der Dialogbox *Voreinstellungen* werden die Indizes für nicht-normalverteilte Daten berechnet, indem zuerst analoge z-Werte für die jeweilige Verteilung bestimmt werden. Bei einer Normalverteilung misst der z-Wert die Anzahl der Standardabweichungen vom Prozessmittelwert bis zu einer Spezifikationsgrenze und steht im direkten Zusammenhang zur Wahrscheinlichkeit, dass eine Beobachtung jenseits dieser Grenze liegt. Bei einer Nicht-Normalverteilung wird ein äquivalenter z-Wert berechnet, indem zunächst die Wahrscheinlichkeit bestimmt wird, die Grenzen zu überschreiten, und anschließend der z-

Wert zu dieser Wahrscheinlichkeit. Nach der Berechnung äquivalenter z-Werte sowohl für die obere als auch die untere Spezifikationsgrenze kann C_{pk} bestimmt werden aus:

$$C_{pk} = \min(Z_{usg}, Z_{osg})/3$$

> ANMERKUNG: Obwohl die Dialogbox *Voreinstellungen* die Möglichkeit der Berechnung der Fähigkeitsindizes aus den Perzentilen anstelle der äquivalenten z-Werte enthält, würde diese Möglichkeit die gebräuchliche Beziehung zwischen den Fähigkeitsindizes und DPM zerstören.

15.5 Six-Sigma-Rechner

Der Index C_{pk} ist eine wertvolle Zusammenfassung der Prozessfähigkeit. Vorausgesetzt, er kann richtig berechnet werden, steht er in Beziehung zu DPM. Das Menü *Werkzeuge* in STATGRAPHICS Centurion XVI enthält einen *Six-Sigma-Rechner*, der zwischen den beiden umrechnet, vorausgesetzt, dass entweder:

1. die Daten normalverteilt sind oder
2. äquivalente z-Werte zur Berechnung der Indizes verwendet werden.

Die Dateneingabe-Dialogbox des *Six-Sigma-Rechners* wird hier gezeigt:

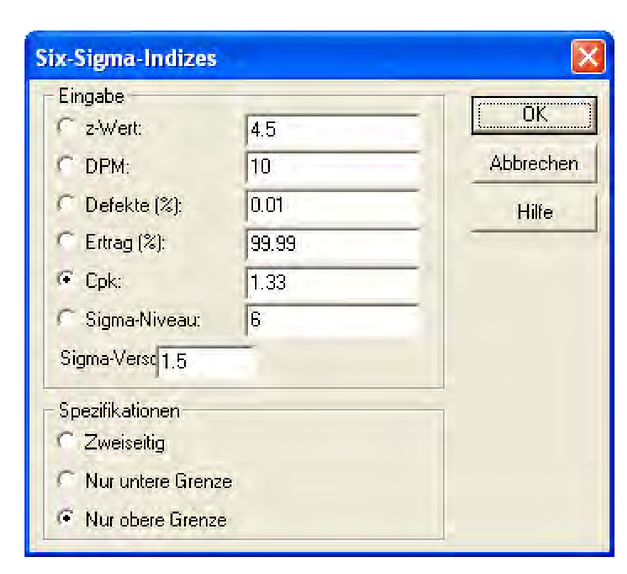


Abbildung 15.17: Six-Sigma-Rechner

Zur Ausführung der Prozeduren:

1. Wählen Sie eins der Optionsfelder und geben Sie einen Wert für die entsprechende Statistik ein.

2. Wenn Sie die Werte nur auf Basis der nächstgelegenen Spezifikationsgrenze berechnen möchten, markieren Sie entweder das Optionsfeld *Nur untere Grenze* oder *Nur obere Grenze*.

3. Geben Sie den Wert ein, den Sie als Langzeit-Veränderung im Prozessmittelwert vermuten. In der Six-Sigma-Methodik wird häufig angenommen, dass der Prozessmittelwert um seinen Langzeitwert mit 1,5 Sigma pendelt.

4. Drücken Sie die Schaltfläche *Berechnen*, um den entsprechenden Wert der anderen Statistik anzuzeigen.



Abbildung 15.18: Äquivalente Werte der Qualitätsindizes

Angenommen, der Prozess verändert sich nicht, dann entspricht ein C_{pk}–Wert von 1,33 etwa 33 DPM jenseits der nächstgelegenen Spezifikationsgrenze.

Kapitel

16

Tutorial 7: Versuchsplanung

Planen von Experimenten als Unterstützung in der Prozessentwicklung und -optimierung.

Nicht alle Daten sind gleich guter Qualität. Häufig liefert eine kleine, aber gut geplante Studie mehr Informationen als eine große, aber schlecht entworfene Untersuchung. Dieses abschließende Tutorial beschreibt einige der Möglichkeiten, die STATGRAPHICS Centurion XVI für das Entwerfen und Analysieren geplanter Experimente bietet.

Betrachten wir den Fall eines Ingenieurs, der untersuchen möchte, welche von mehreren Prozessvariablen den größten Einfluss auf ein Endprodukt haben. Er beabsichtigt, die Auswirkung der Änderung von fünf Faktoren zu ermitteln: Eingangstemperatur, Fließgeschwindigkeit, Konzentration, Umwälzungsrate und Katalysator in Prozent. In der Praxis könnte man auf verschiedene Weise an dieses Problem herangehen:

1. *Versuch und Irrtum*: willkürliche Auswahl verschiedener Faktorkombinationen jedes Mal, wenn ein Experiment durchgeführt wird. Diese Herangehensweise liefert selten nutzbringende Informationen.

2. *Experimentieren mit jeweils einem Faktor*: alle Faktoren bis auf einen konstant lassen, um den Effekt dieses einen Faktors zu bestimmen. Diese Herangehensweise ist extrem ineffizient und kann in die Irre führen, falls es irgendwelche Wechselwirkungen der Faktoren gibt.

3. *Verwenden eines statistischen Versuchsplanes*: eine Reihe von Experimenten gezielt planen, welche die meisten Informationen über die Faktoren und deren Wechselwirkungen in nur wenigen Versuchen liefern.

In diesem Tutorial wird beschrieben, wie man einen Versuchsplan unter Anwendung der dritten Variante erstellt und wie die Ergebnisse analysiert werden können.

16.1 Entwerfen eines Versuchsplanes

STATGRAPHICS Centurion XVI besitzt einen Versuchsplanungs-Assistenten (DoE-Assistent), der den Anwender durch den Prozess des Entwerfens und Auswertens von geplanten Experimenten führt. So rufen Sie den DoE-Assistenten auf:

1. Wenn Sie das klassische Menü nutzen, wählen Sie *DOE – DoE-Assistent*.

2. Wenn Sie das Six-Sigma-Menü nutzen, wählen Sie *Verbessern – DoE-Assistent*.

Jetzt wird ein neues Fenster mit einer Symbolleiste präsentiert, wo Sie durch eine Folge von 12 Schritten geführt werden:

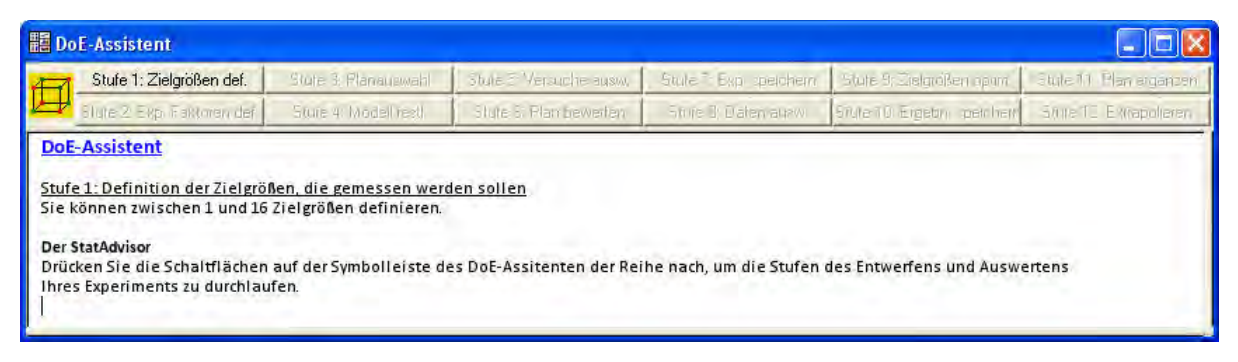


Abbildung 16.1: Hauptfenster des DoE-Assistenten mit einer 12-Stufen-Symbolleiste

Mit den ersten 7 Schritten wird der Versuchsplan konstruiert, diese werden also durchlaufen, bevor die Experimente stattfinden. Die letzten 5 Schritte werden nach der Durchführung der Versuche ausgeführt und beschäftigen sich mit der Auswertung der Versuchsergebnisse.

Stufe 1: Zielgrößen definieren

Der erste Schritt beim Entwerfen eines Versuchsplanes ist die Definition der Zielgrößen, die während jedes einzelnen Versuches gemessen werden. Nach dem Drücken der Schaltfläche *Stufe 1* erscheint folgende Dialogbox:

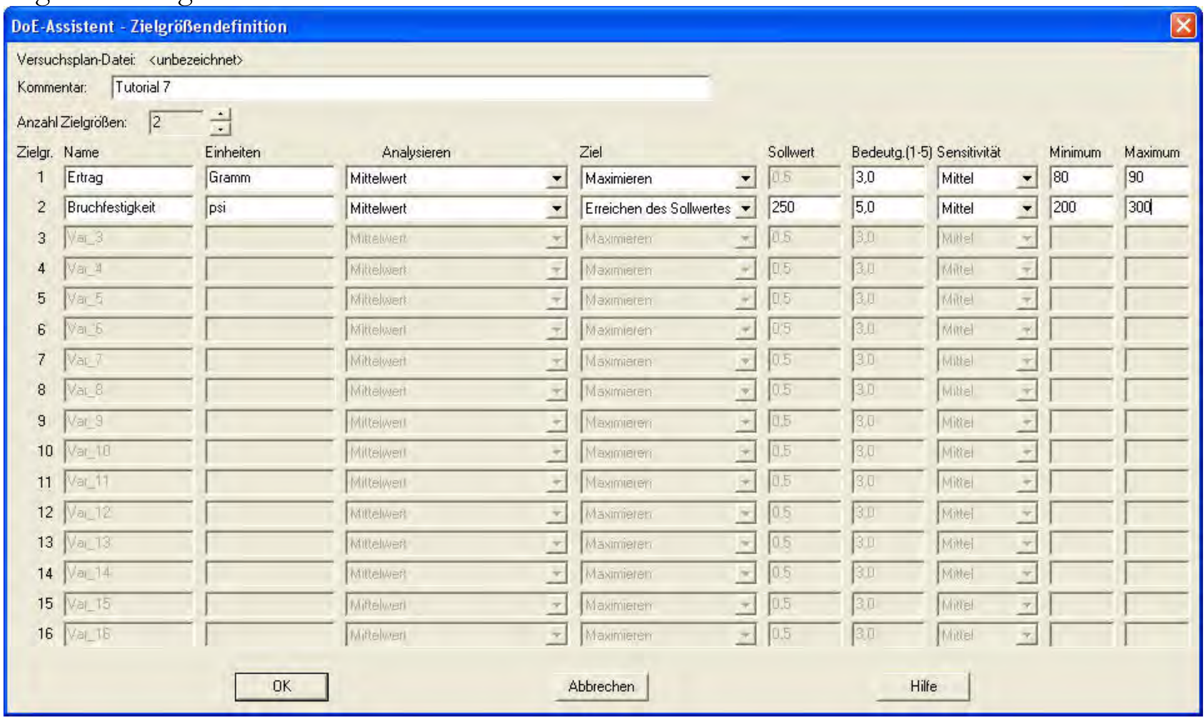


Abbildung 16.2: Definition der Zielgrößen

In unserem Beispiel gibt es zwei Zielgrößen: *Ertrag* in Gramm und *Bruchfestigkeit* in psi (pounds per square inch). Das Ziel des Experiments besteht in der Maximierung von *Ertrag*, während die *Festigkeit* möglichst nahe bei 250 liegen soll. Die vier rechts liegenden Spalten dienen zur Ausgleichung der Anforderungen an die beiden Zielgrößen, falls diese sich widersprechen. *Bedeutung* legt die Wichtigkeit jeder Zielgröße auf einer Skala von 1 bis 5 fest, wobei 5 der größten Bedeutung zugeordnet ist. Die *Minimum*- und *Maximum*-Werte beschreiben den gewünschten Bereich für jede Zielgröße, wobei *Sensitivität* anzeigt, wie wichtig es ist, der besten Position in diesem Bereich möglichst nahe zu kommen. In unserem Beispiel ist die *Festigkeit* wichtiger als der *Ertrag* und hat deshalb einen höheren Wert bei Bedeutung. Die Sensitivität ist für beide Zielgrößen auf „Mittel" eingestellt, d.h. die Erwünschtheit für jede Zielgröße steigt innerhalb des angegebenen Bereiches in linearer Form an.

Stufe 2: Definition der experimentellen Faktoren

Die Schaltfläche *Stufe 2* dient zur Eingabe der Informationen über die experimentellen Faktoren, die während des Experiments variiert werden sollen. Folgende Dialogbox wird angezeigt:

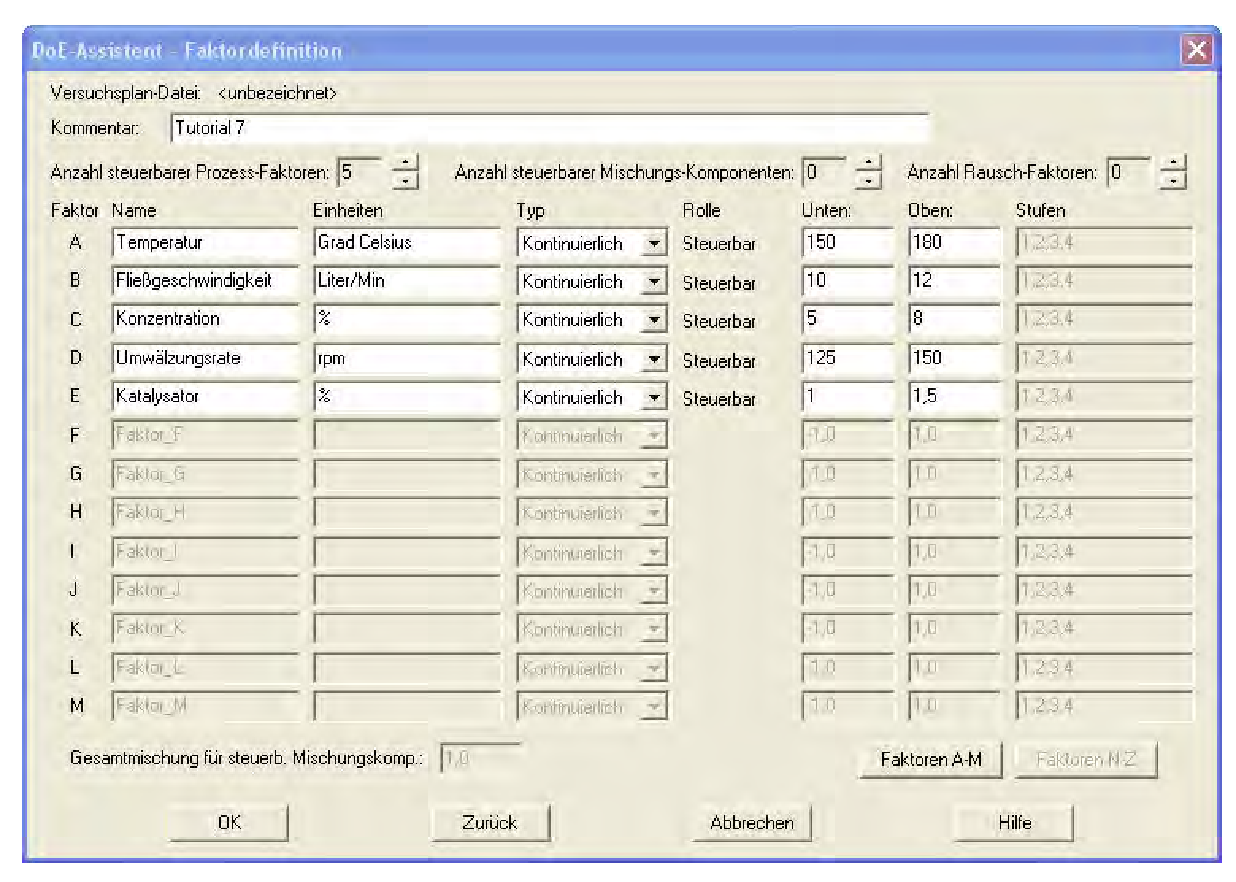


Abbildung 16.3: Definition der experimentellen Faktoren

Im Beispiel sind fünf steuerbare Prozessfaktoren zu variieren. Geben Sie für jeden Faktor den Namen, seine Einheiten und den Bereich ein, über den variiert werden soll. Alle Faktoren hier sind *kontinuierlich*, sie können auf jeden beliebigen Wert zwischen der unteren und der oberen Stufe eingestellt werden.

Stufe 3: Auswahl des Versuchsplanes

Der dritte Schritt beim Entwerfen ist die Auswahl des Plantyps, welcher ausgeführt werden soll. Nach Drücken der Schaltfläche *Stufe 3* erscheint folgende Dialogbox:

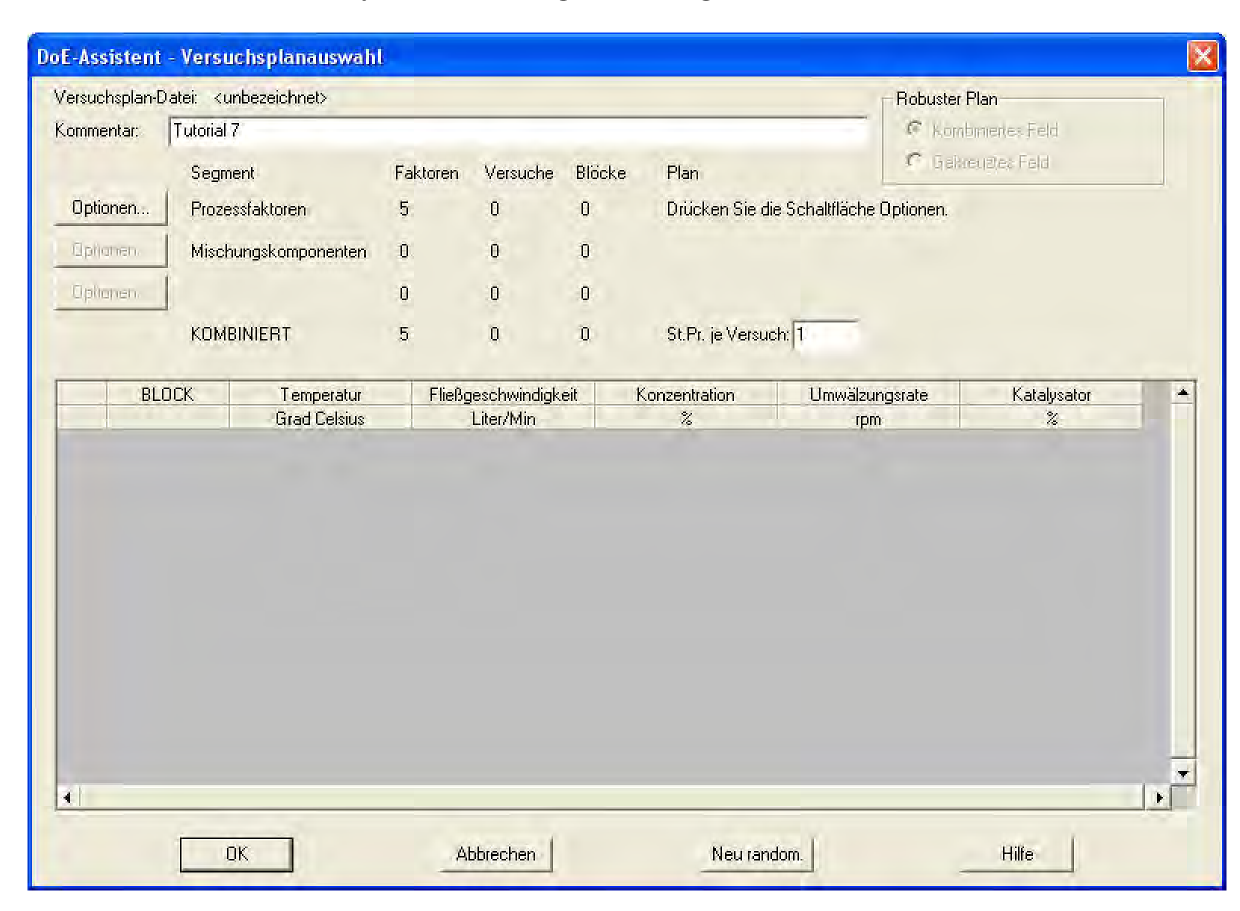


Abbildung 16.4: Dialogbox für Planauswahl

Um einen Versuchsplan für die 5 Faktoren zu wählen, drücken Sie die Schaltfläche *Optionen*. Eine Auswahl von Versuchsplan-Klassen, die für 5 kontinuierliche Faktoren geeignet sind, wird angezeigt:

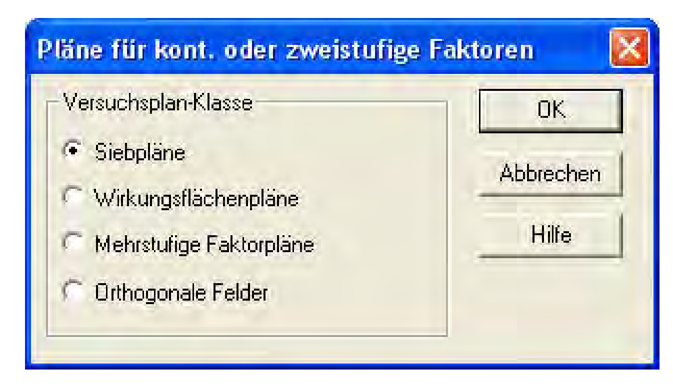


Abbildung 16.5: Dialogbox mit den verfügbaren Versuchsplan-Klassen

Da wir einen Siebplan (Versuchsplan erster Ordnung) verwenden wollen, wird einfach *OK* gedrückt.
Die nächste Dialogbox dient zur Auswahl des gewünschten Plans aus einer Reihe von Siebplänen, die für 5 Faktoren geeignet sind:

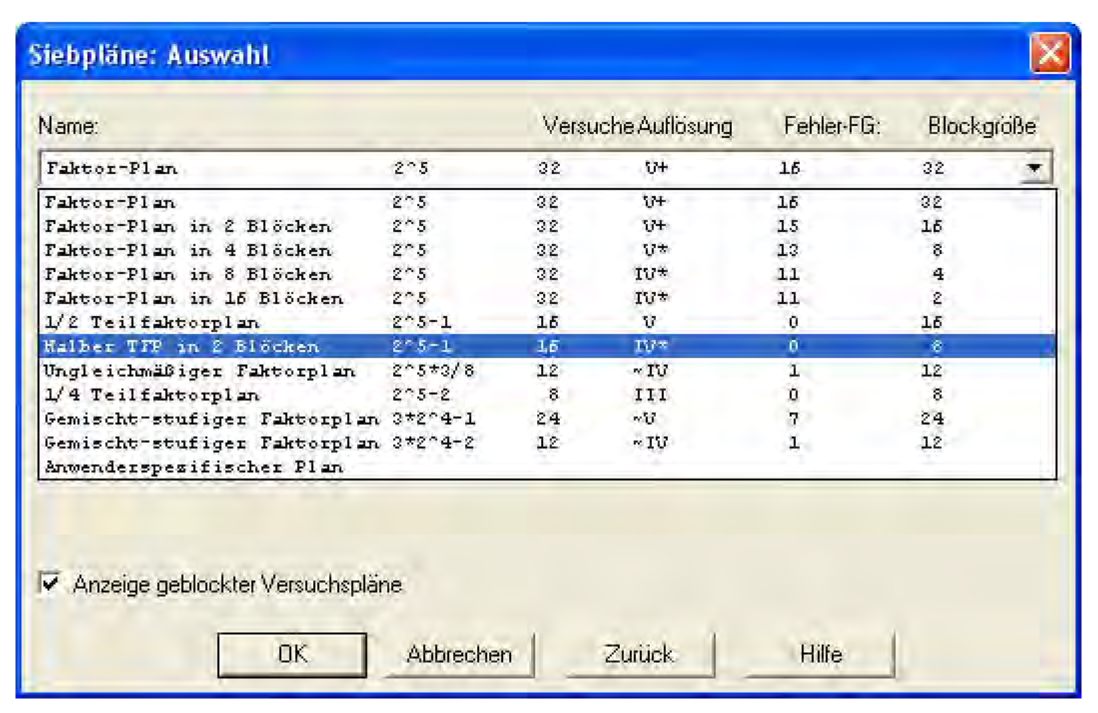


Abbildung 16.6: Versuchsplan-Liste

Um diese Liste von verfügbaren Siebplänen für 5 Faktoren zu sehen, müssen Sie auf den Pfeil rechts klicken. Es werden angezeigt:

1. ***Name***: Name des verfügbaren Versuchsplanes.
2. ***Versuche***: Anzahl von Einzelversuchen im Basisplan, bevor irgendwelche Zentrumsversuche oder Wiederholungen von Versuchen hinzugefügt werden.
3. ***Auflösung***: die Auflösung des Versuchsplanes. Mit einem Plan der Auflösung V kann man alle Haupteffekte und Zweifaktor-Wechselwirkungen schätzen. Mit einem Plan der Auflösung IV kann man alle Haupteffekte schätzen, die Zweifaktor-Wechselwirkungen können aber untereinander oder mit dem Blockfaktor vermengt sein. Bei Auflösung III sind Zweifaktor-Wechselwirkungen mit Haupteffekten vermengt.
4. ***Fehler-FG.***: die Anzahl von Freiheitsgraden, die für die Schätzung des Versuchsfehlers verfügbar sind. Die Power eines statistischen Tests steht in Beziehung zu den Freiheitsgraden, ebenso wie die Gesamtanzahl der Versuche im Experiment. Im Allgemeinen sollten mindestens drei Freiheitsgrade verfügbar sein, wobei eine größere Anzahl vorzuziehen wäre.
5. ***Blockgröße***: die Anzahl der Versuche im größten Block.

In unserem Fall wurde ein Teilfaktorplan in zwei Blöcken mit jeweils acht Versuchen ausgewählt.

In der letzten Dialogbox werden Zentrumspunkte und komplette Wiederholungen hinzugefügt:



Abbildung 16.7: Optionen für geblockte Siebpläne

Die Eingabefelder legen fest:

1. *Zentrumspunkte*: die Anzahl von Versuchen, die im Zentrum des Versuchsraumes durchgeführt werden sollen. Das Hinzufügen von Zentrumspunkten ist eine gute Möglichkeit, um die Freiheitsgrade für den Versuchsfehler zu erhöhen.

2. *Anordnung*: die Platzierung der Zentrumspunkte. Die einfachsten Varianten sind *Zufällig*, wobei die Zentrumspunktversuche zufällig zwischen den anderen Versuchen verteilt werden, und *Regelmäßige Abstände*, wobei die Zentrumspunktversuche gleichmäßig über den gesamten Versuchsplan verteilt werden.

3. *Komplette Wiederholung*: die Anzahl von zusätzlichen kompletten Wiederholungen aller Einzelversuche. Die Wiederholung des gesamten Planes kann die Anzahl der Versuche sehr schnell anwachsen lassen.

4. *Randomisieren*: gibt an, ob die Versuche in einer zufälligen Reihenfolge abgearbeitet werden sollen. Das sollte möglichst immer gemacht werden, um zu verhindern, dass externe Störgrößen (wie Änderungen im Prozess über die Zeit) die Ergebnisse verzerren.

Für unser aktuelles Experiment werden vier Zentrumspunkte benötigt, womit das endgültige Experiment aus 20 Einzelversuchen besteht. Es wird außerdem gefordert, die Versuchsreihenfolge zufällig zu wählen, d.h. dass die zehn Versuche jedes Blocks zufällig nacheinander ablaufen sollen.

Nach der abschließenden Dialogbox wird das Planauswahl-Fenster mit den auszuführenden experimentellen Versuchen gefüllt:

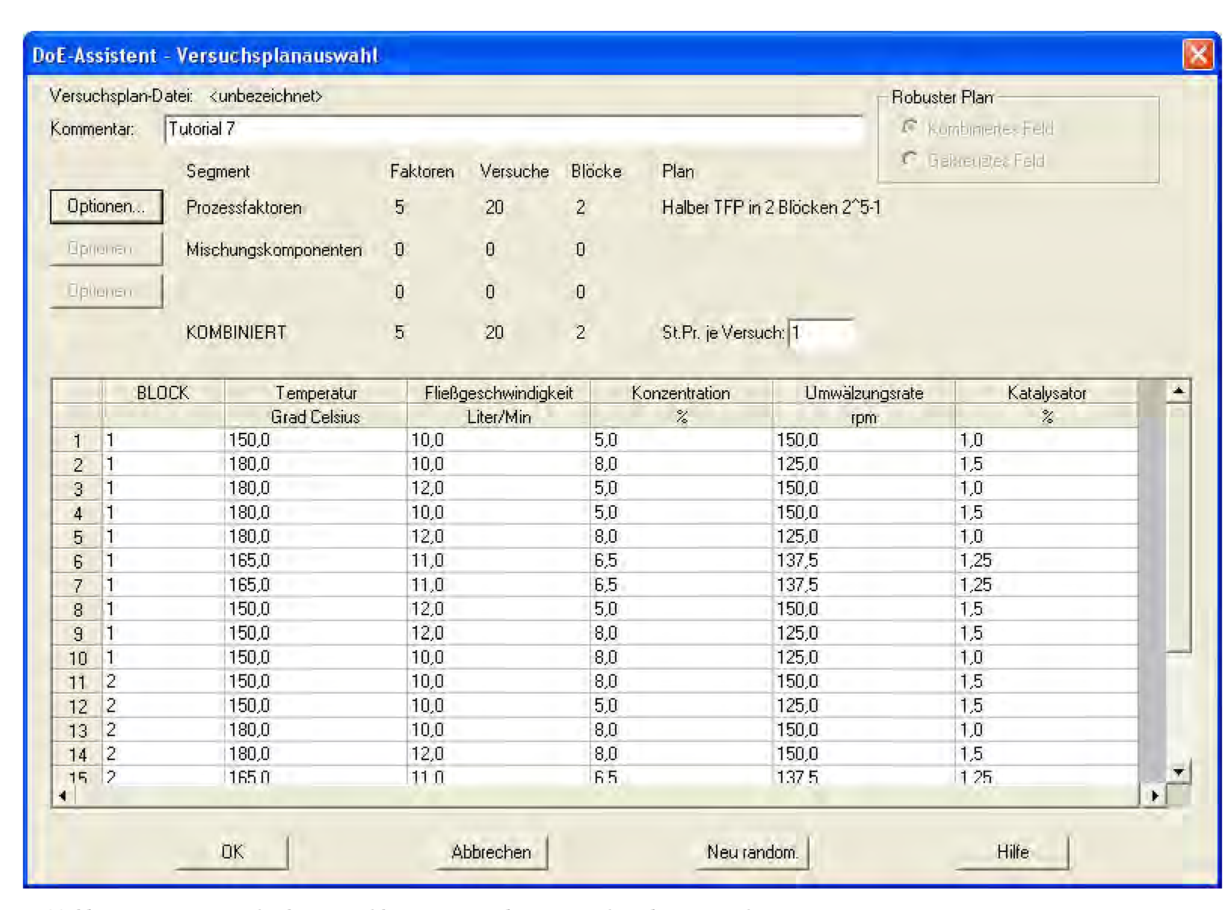
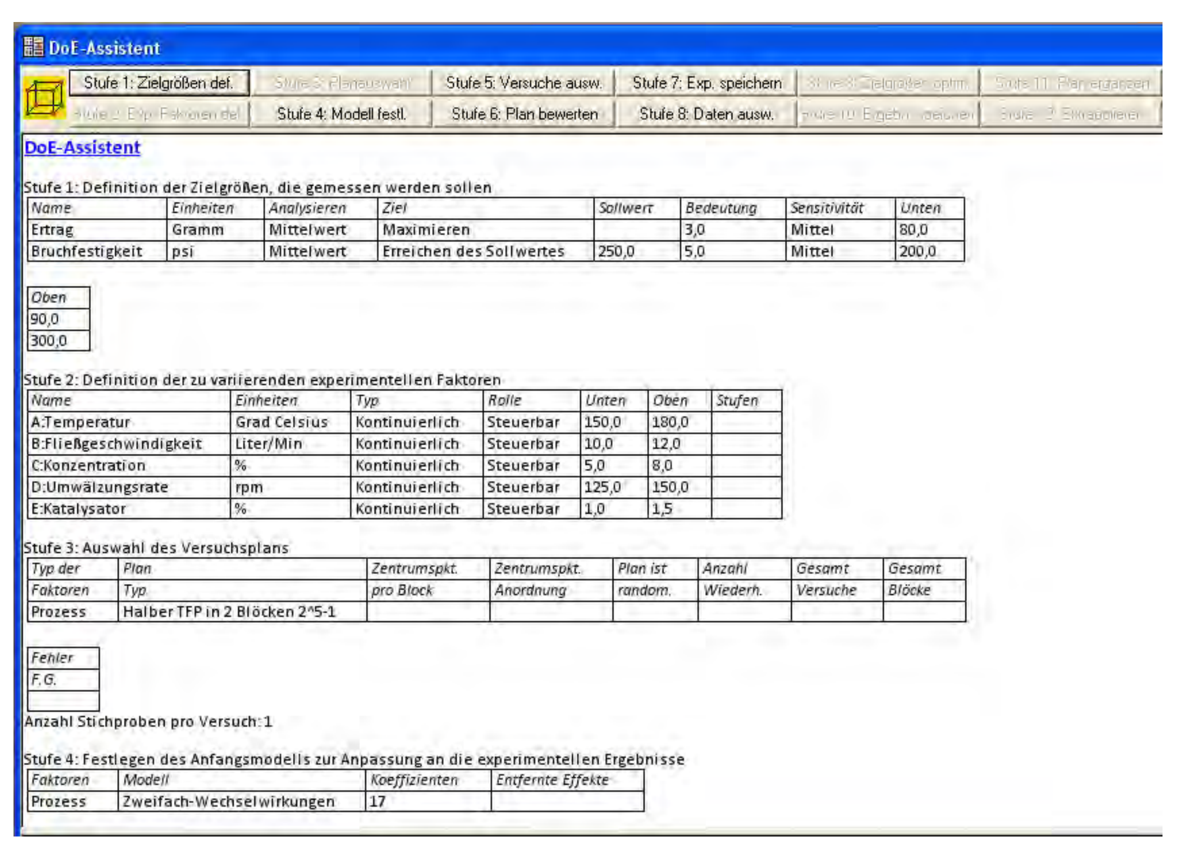


Abbildung 16.8: Versuchsplanauswahl-Fenster mit den auszuführenden Versuchen

Wenn Sie mit diesem Plan einverstanden sind, drücken Sie nochmals *OK*, um zum Fenster des *DoE-Assistenten* zurückzukehren. Dort wird eine Zusammenfassung wie folgt angezeigt:



DoE-Assistent

Stufe 1: Definition der Zielgrößen, die gemessen werden sollen

Name	Einheiten	Analysieren	Ziel	Sollwert	Bedeutung	Sensitivität	Unten
Ertrag	Gramm	Mittelwert	Maximieren		3,0	Mittel	80,0
Bruchfestigkeit	psi	Mittelwert	Erreichen des Sollwertes	250,0	5,0	Mittel	200,0

Oben
90,0
300,0

Stufe 2: Definition der zu variierenden experimentellen Faktoren

Name	Einheiten	Typ	Rolle	Unten	Oben	Stufen
A:Temperatur	Grad Celsius	Kontinuierlich	Steuerbar	150,0	180,0	
B:Fließgeschwindigkeit	Liter/Min	Kontinuierlich	Steuerbar	10,0	12,0	
C:Konzentration	%	Kontinuierlich	Steuerbar	5,0	8,0	
D:Umwälzungsrate	rpm	Kontinuierlich	Steuerbar	125,0	150,0	
E:Katalysator	%	Kontinuierlich	Steuerbar	1,0	1,5	

Stufe 3: Auswahl des Versuchsplans

Typ der Faktoren	Plan Typ	Zentrumspkt. pro Block	Zentrumspkt. Anordnung	Plan ist random.	Anzahl Wiederh.	Gesamt Versuche	Gesamt Blöcke
Prozess	Halber TFP in 2 Blöcken 2^5-1						

Fehler F.G.

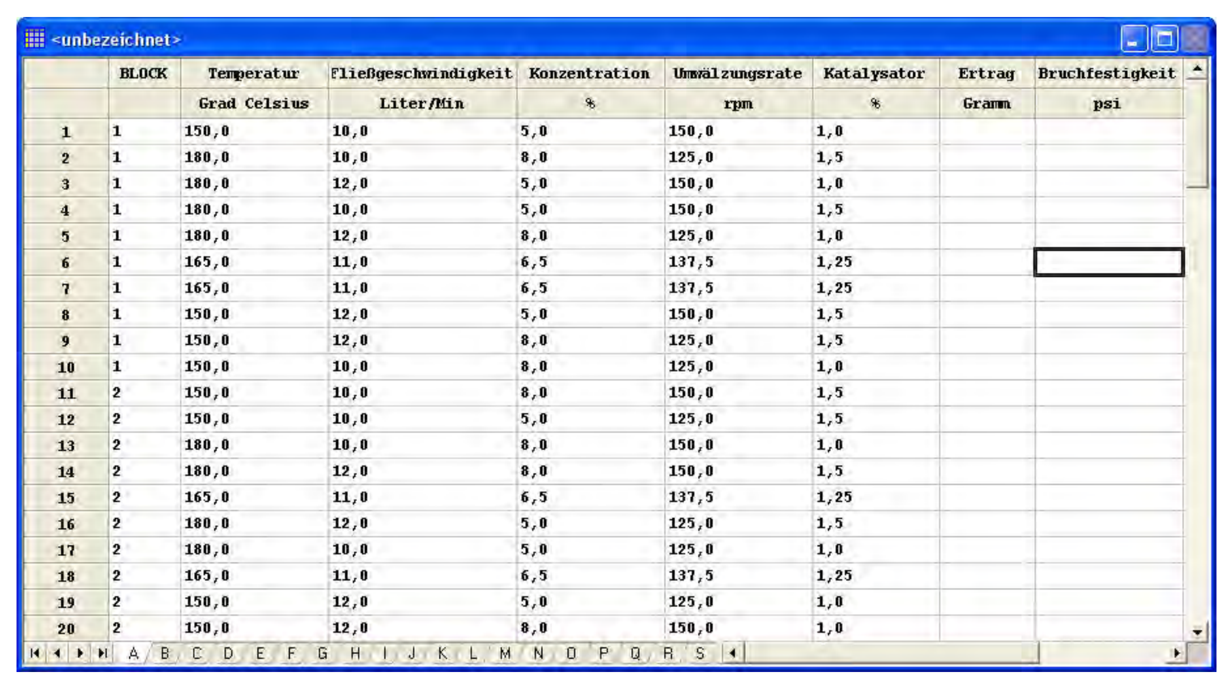
Anzahl Stichproben pro Versuch: 1

Stufe 4: Festlegen des Anfangsmodells zur Anpassung an die experimentellen Ergebnisse

Faktoren	Modell	Koeffizienten	Entfernte Effekte
Prozess	Zweifach-Wechselwirkungen	17	

Abbildung 16.9: DoE-Assistent nach der Auswahl eines Versuchsplanes

Gleichzeitig wird der Plan in das Tabellenblatt A des Datenbuches von STATGRAPHICS Centurion XVI geladen:



<unbezeichnet>

	BLOCK	Temperatur	Fließgeschwindigkeit	Konzentration	Umwälzungsrate	Katalysator	Ertrag	Bruchfestigkeit
		Grad Celsius	Liter/Min	%	rpm	%	Gramm	psi
1	1	150,0	10,0	5,0	150,0	1,0		
2	1	180,0	10,0	8,0	125,0	1,5		
3	1	180,0	12,0	5,0	150,0	1,0		
4	1	180,0	10,0	5,0	150,0	1,5		
5	1	180,0	12,0	8,0	125,0	1,0		
6	1	165,0	11,0	6,5	137,5	1,25		
7	1	165,0	11,0	6,5	137,5	1,25		
8	1	150,0	12,0	5,0	150,0	1,5		
9	1	150,0	12,0	8,0	125,0	1,5		
10	1	150,0	10,0	8,0	125,0	1,0		
11	2	150,0	10,0	8,0	150,0	1,5		
12	2	150,0	10,0	5,0	125,0	1,5		
13	2	180,0	10,0	8,0	150,0	1,0		
14	2	180,0	12,0	8,0	150,0	1,5		
15	2	165,0	11,0	6,5	137,5	1,25		
16	2	180,0	12,0	5,0	125,0	1,5		
17	2	180,0	10,0	5,0	125,0	1,0		
18	2	165,0	11,0	6,5	137,5	1,25		
19	2	150,0	12,0	5,0	125,0	1,0		
20	2	150,0	12,0	8,0	150,0	1,0		

Abbildung 16.10: Fertiger Plan

Das Datenblatt enthält eine Spalte mit den Blocknummern, fünf Spalten mit den Einstellungen für die Faktoren und zwei Spalten, die für die Zielgrößenwerte vorgesehen sind, nachdem die Versuche durchgeführt wurden.

Stufe 4: Modell festlegen

Der DoE-Assistent wird sich bei der Auswertung des erzeugten Versuchsplanes auf das hier festgelegte Modell beziehen. Nach dem Drücken der Schaltfläche *Stufe 4* wird folgende Dialogbox angezeigt:

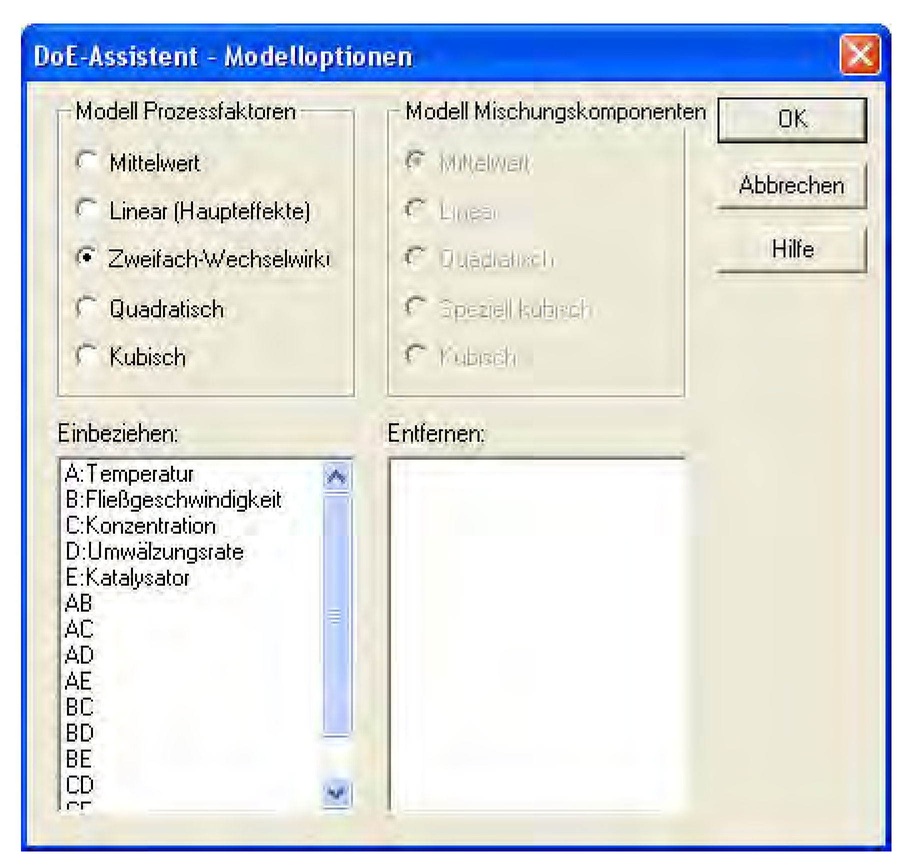


Abbildung 16.11: Dialogbox zur Modellauswahl

Sie sollten das komplexeste Modell auswählen, das Sie für die Daten verwenden können. Im Fall des zweistufigen Faktorplanes ist das komplexeste Modell für die Anpassung eines mit Zweifachwechselwirkungen, das folgendermaßen definiert ist:

$$Y = \begin{array}{l} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{14} x_1 x_4 \\ + \beta_{15} x_1 x_5 + \beta_{23} x_2 x_3 + \beta_{24} x_2 x_4 + \beta_{25} x_2 x_5 + \beta_{34} x_3 x_4 + \beta_{35} x_3 x_5 + \beta_{45} x_4 x_5 \end{array}$$

Es enthält jeden experimentellen Faktor selbst (Haupteffekte) sowie Glieder mit allen Paaren von Faktoren (Zweifaktor-Wechselwirkungen). Einzelne Glieder können aus dem Modell entfernt werden, indem man in der Dialogbox doppelt auf deren Namen klickt, wodurch sie in das *Entfernen*-Feld der Dialogbox verschoben werden. In unserem Fall wollen wir das komplette Zweifaktor-Wechselwirkungsmodell auswählen.

Stufe 5: Versuche auswählen

Bei komplexeren Versuchsplänen kann es wünschenswert sein, nur eine Teilmenge der im Schritt 3 kreierten Versuche auszuführen. Nach dem Drücken der Schaltfläche *Stufe 5* kann ein Auswahlalgorithmus verwendet werden, der eine *D-optimale* Teilmenge von Versuchen heraussucht. In unserem Fall sollen alle Versuche ausgeführt werden, so dass wir Stufe 5 überspringen

Stufe 6: Plan bewerten

Nach dem Drücken der Schaltfläche *Stufe 6* wird eine Dialogbox angezeigt, die eine Anzahl von Tabellen und Grafiken enthält, welche zum Fenster des DoE-Assistenten hinzugefügt werden können:

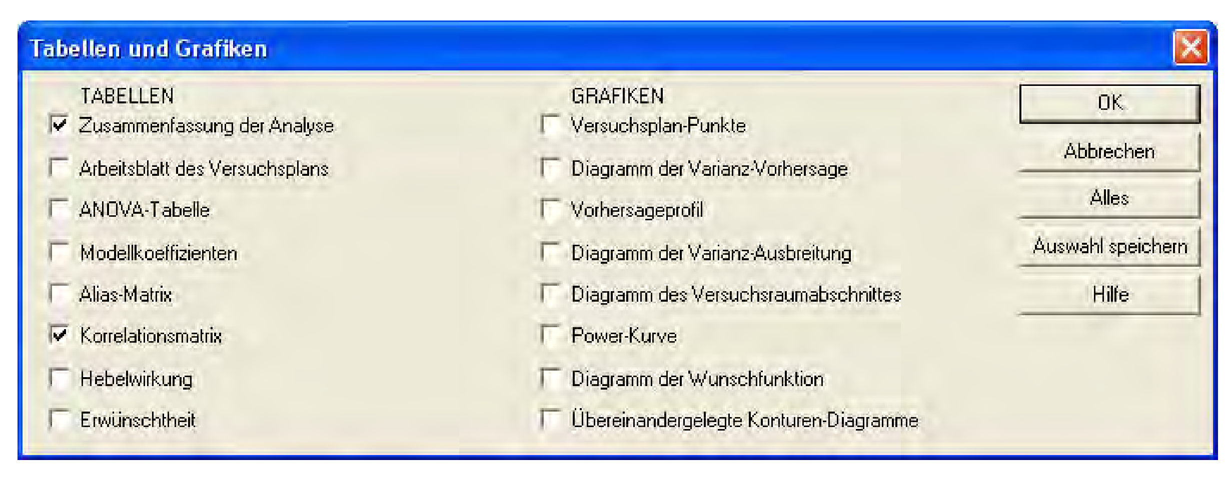


Abbildung 16.12: Tabellen und Grafiken für die Bewertung des gewählten Versuchsplanes

Eine sinnvolle Option für Siebpläne ist die *Korrelationsmatrix*, die anzeigt, ob irgendwelche Vermengungen unter den Gliedern des anzupassenden Modells existieren:

Korrelationsmatrix

	Block	A	B	C	D	E	AB	AC	AD	AE	BC	BD	BE	CD
Block	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,8944
A	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
B	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
C	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
D	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
E	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
AB	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
AC	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
AD	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000	0,0000
AE	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000	0,0000
BC	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000	0,0000
BD	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000	0,0000
BE	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000	0,0000
CD	0,8944	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,0000
CE	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DE	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

	CE	DE
Block	0,0000	0,0000
A	0,0000	0,0000
B	0,0000	0,0000
C	0,0000	0,0000
D	0,0000	0,0000
E	0,0000	0,0000
AB	0,0000	0,0000
AC	0,0000	0,0000

Abbildung 16.13: Korrelationsmatrix für den gewählten Versuchsplan

Ein Wert ungleich Null in irgendeiner Zelle außerhalb der Hauptdiagonalen bedeutet, dass die Effekte dieser Zeile und Spalte vermengt sind und nicht sauber getrennt werden können. Im aktuellen Plan hat die Wechselwirkung CD eine hohe Korrelation mit den Blöcken. Beachten Sie, dass die Möglichkeit zum Schätzen der Wechselwirkung zwischen C und D in diesem Plan willkürlich geopfert wurde (hier: C - *Konzentration*, D - *Umwälzungsrate*). Sofern das eine für Sie wesentliche Wechselwirkung sein sollte, können Sie die Reihenfolge der Variablen verändern, so dass C und D anderen Variablen entsprechen, die sich vielleicht nicht gegenseitig beeinflussen.

Stufe 7: Experiment speichern

Nach dem Drücken der Schaltfläche *Stufe 7* kann das Experiment in einer Datei gespeichert werden. Es wird folgende Dialogbox angezeigt:

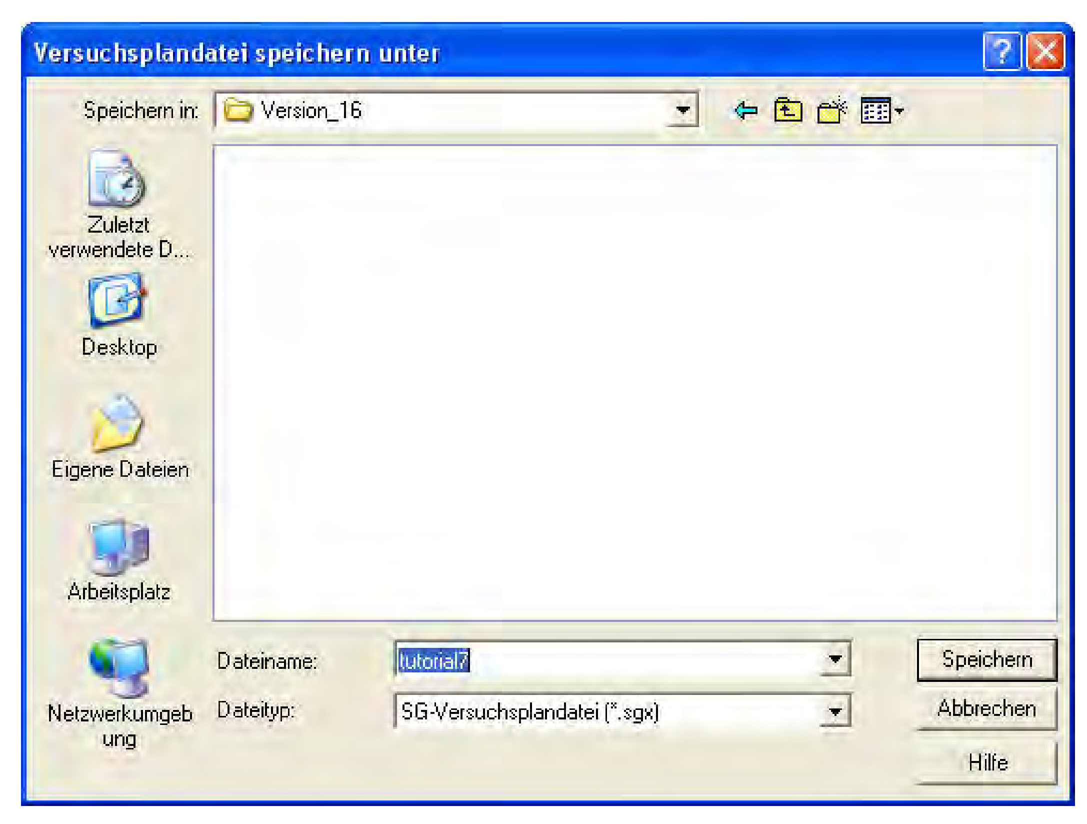


Abbildung 16.14: Dialogbox zum Speichern des Experiments

Versuchspläne, die mit dem DoE-Assistenten erzeugt wurden, werden in Dateien mit der Endung *.sgx* gespeichert. Diese sind ähnlich den normalen Datendateien mit der Besonderheit, dass sie zusätzliche Informationen zum Versuchsplan und zum gewählten statistischen Modell enthalten.

16.2 Auswerten der Ergebnisse

Nach der Erstellung des Versuchsplanes führt unser Ingenieur die angegebenen 20 Versuche aus. Dann startet er erneut das Programm und gibt die Messwerte für *Ertrag* und *Festigkeit* in das Datenblatt für das Experiment ein. Um die Auswertung nachzuvollziehen, können Sie die Datei *tutorial7.sgx* in der gleichen Weise wie eine normale STATGRAPHICS-Datendatei laden, indem Sie im Hauptmenü *Datei – Öffnen – Datenquelle öffnen* aufrufen (Bem.: Diese Datei wird hier im Folgenden mit deutschen Spaltenbezeichnungen verwendet).
Beim Öffnen einer Versuchsplandatei wird automatisch auch das Hauptfenster des DoE-Assistenten geöffnet.

Stufe 8: Daten auswerten

Um die experimentell ermittelten Daten auszuwerten, drücken Sie die Schaltfläche *Stufe 8*. Die Analyse startet mit der Anzeige folgender Dialogbox:

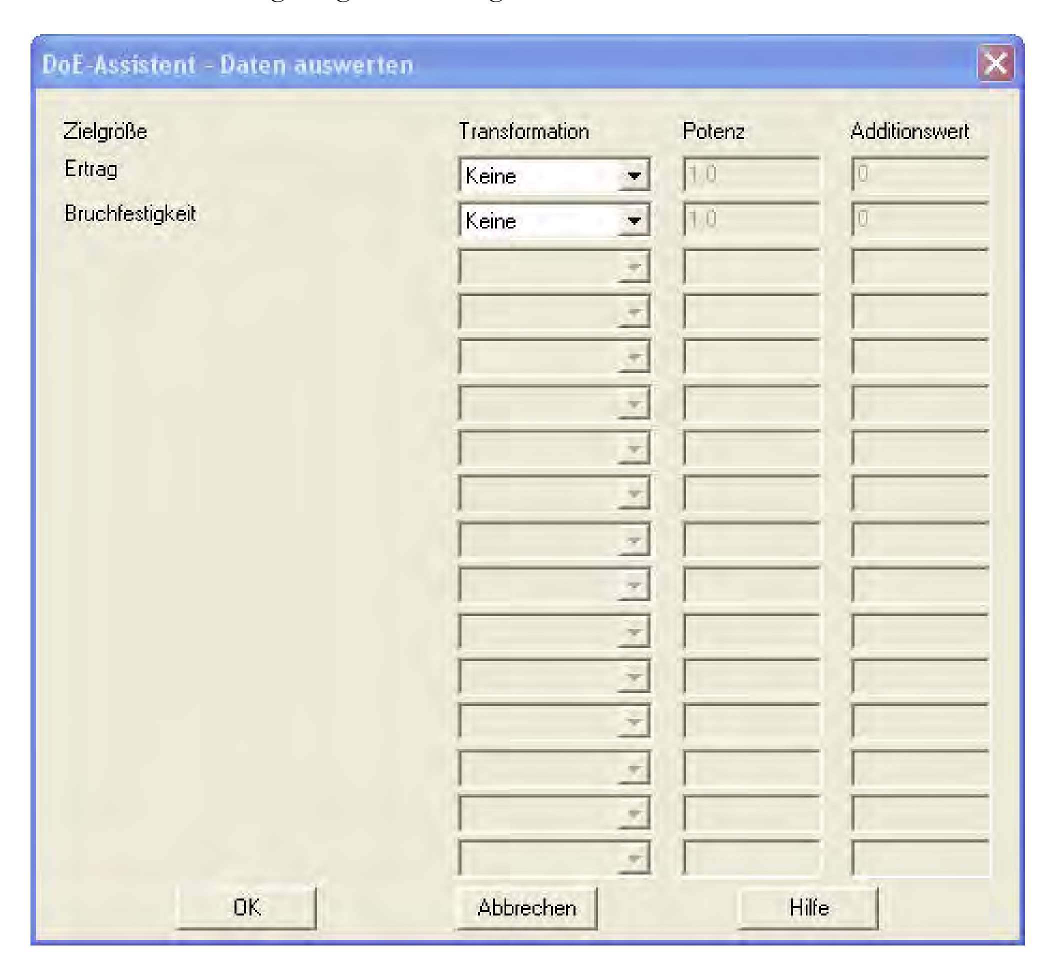
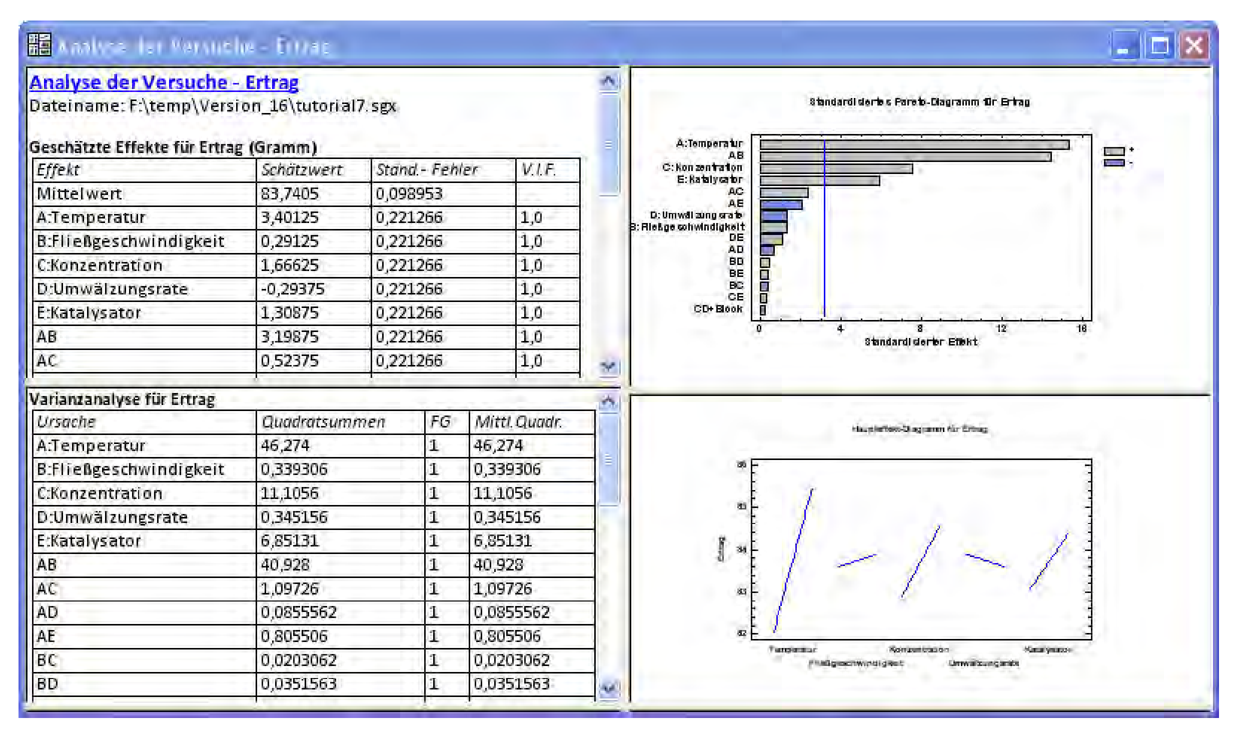


Abbildung 16.15: Dateneingabe-Dialogbox für die Auswertung des Versuchsplanes

Sofern erforderlich, kann man für jede Variable eine Transformation festlegen. Da es sich hier um kontinuierliche Zielgrößen handelt, können solche Transformationen notwendig werden, wenn die Varianz der Zielgröße mit dem Mittelwert ansteigt. In unserem Fall ist das aber nicht erforderlich.

Nach dem Drücken der *OK*-Schaltfläche wird für jede Zielgröße ein Analysefenster geöffnet.

Das Analyse-Fenster für *Ertrag* zeigt anfangs folgende Ergebnisse:



Effekt	Schätzwert	Stand.- Fehler	V.I.F.
Mittelwert	83,7405	0,098953	
A:Temperatur	3,40125	0,221266	1,0
B:Fließgeschwindigkeit	0,29125	0,221266	1,0
C:Konzentration	1,66625	0,221266	1,0
D:Umwälzungsrate	-0,29375	0,221266	1,0
E:Katalysator	1,30875	0,221266	1,0
AB	3,19875	0,221266	1,0
AC	0,52375	0,221266	1,0

Ursache	Quadratsummen	FG	Mittl. Quadr.
A:Temperatur	46,274	1	46,274
B:Fließgeschwindigkeit	0,339306	1	0,339306
C:Konzentration	11,1056	1	11,1056
D:Umwälzungsrate	0,345156	1	0,345156
E:Katalysator	6,85131	1	6,85131
AB	40,928	1	40,928
AC	1,09726	1	1,09726
AD	0,0855562	1	0,0855562
AE	0,805506	1	0,805506
BC	0,0203062	1	0,0203062
BD	0,0351563	1	0,0351563

Abbildung 16.16: Analyse-Fenster für die Auswertung der Zielgröße Ertrag

Das Fenster enthält vier Ergebnisfenster:

1. *Zusammenfassung der Analyse*: listet die geschätzten Haupt- und Wechselwirkungseffekte auf.

2. *ANOVA-Tabelle*: enthält *p*-Werte, welche für jeden Effekt anzeigen, ob dieser signifikant ist.

3. *Standardisiertes Pareto-Diagramm*: zeigt die Effekte der Größe nach in absteigender Reihenfolge mit einer Linie, deren Überschreiten statistische Signifikanz bedeutet.

4. *Haupteffekt-Diagramm*: stellt die geschätzte Änderung der Zielgröße dar, wenn sich jeder der Faktoren von der unteren zur oberen Stufe bewegt.

Das standardisierte Pareto-Diagramm in der rechten oberen Ecke kann man verwenden, um sehr schnell festzustellen, welche Effekte am wichtigsten sind:

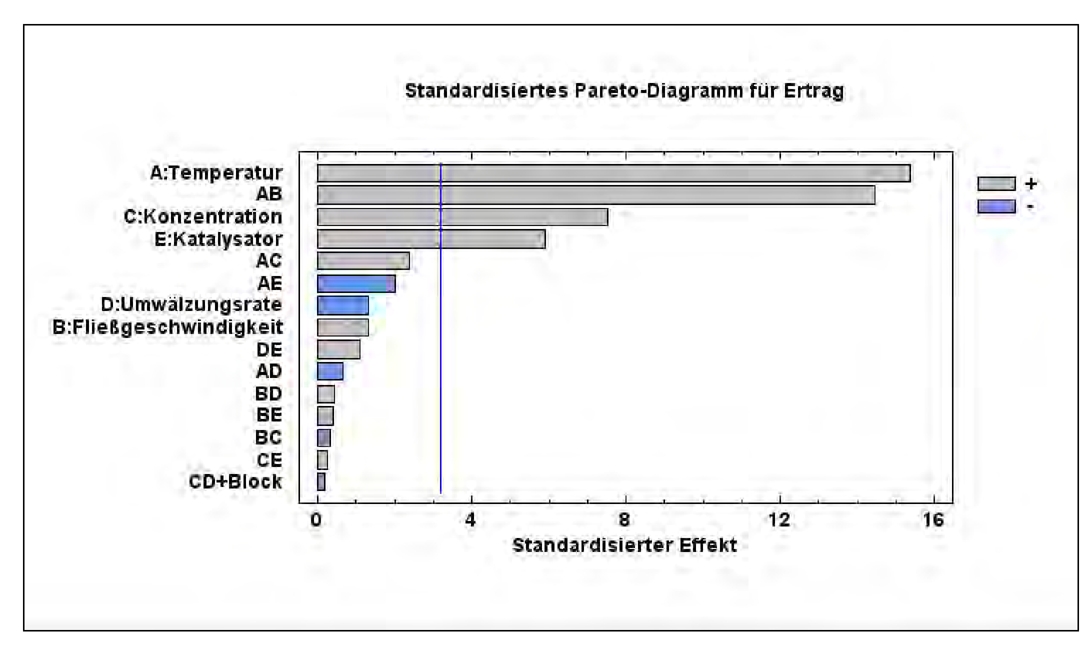


Abbildung 16.17: Standardisiertes Pareto-Diagramm

Die Länge jedes Balkens ist proportional zu einem Wert der t-Statistik, der für den entsprechenden Effekt berechnet wurde. Jeder Effekt mit einem Balken jenseits der vertikalen Linie ist statistisch signifikant mit der gewählten Irrtumswahrscheinlichkeit, die standardmäßig auf 5 % gesetzt ist. In unserem Fall gibt es drei signifikante Haupteffekte: *Temperatur*, *Konzentration* und *Katalysator*. Außerdem gibt es eine signifikante Wechselwirkung zwischen *Temperatur* und *Fließgeschwindigkeit*.

Das *Haupteffekt-Diagramm* im rechten unteren Ergebnisfenster zeigt, wie jeder Faktor den *Ertrag* beeinflusst:

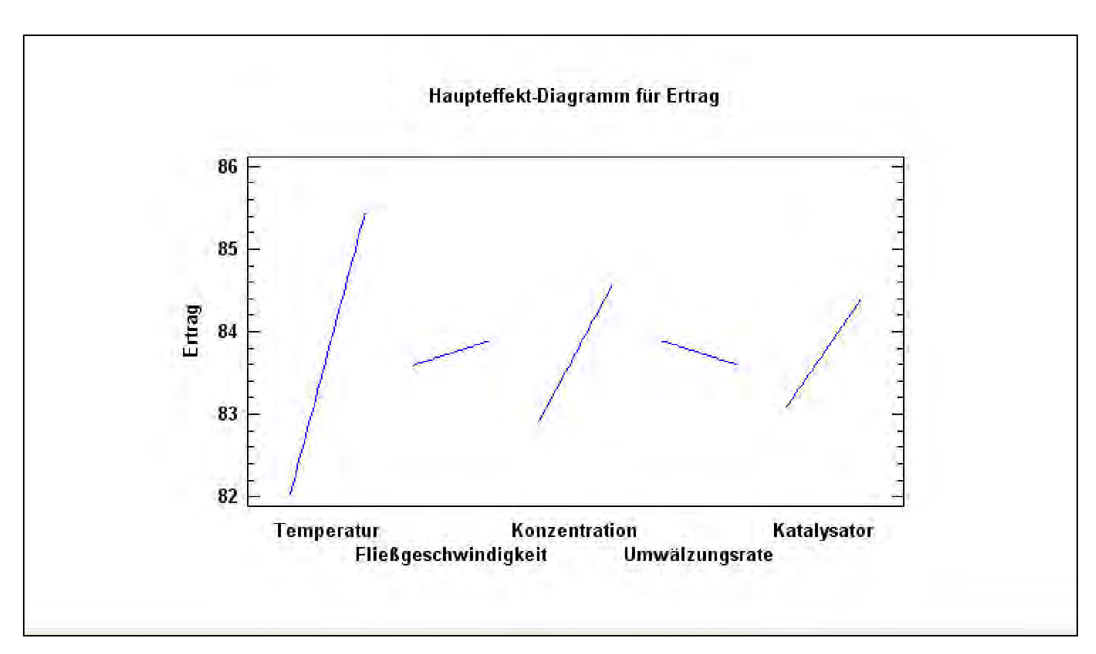


Abbildung 16.18: Haupteffekt-Diagramm

Die Linie gibt die geschätzte Änderung für *Ertrag* an, wenn sich jeder Faktor von der unteren zur oberen Stufe bewegt, wobei die anderen Faktoren auf dem mittleren Wert zwischen deren unterer und oberer Stufe konstant gehalten werden. Beachten Sie, dass die drei Faktoren mit signifikantem Haupteffekt einen größeren Einfluss auf die Zielgröße haben als die anderen. Zum Beispiel ist der mittlere Ertrag für die untere Temperatur ungefähr 82 und für die obere Temperatur etwa 85,4. Die Differenz von 3,4 wird „Haupteffekt" des Faktors *Temperatur* genannt.

Um die Wechselwirkungen zwischen *Temperatur* und *Fließgeschwindigkeit* darzustellen, ruft man zuerst das *Wechselwirkungs-Diagramm* in der Dialogbox *Tabellen und Grafiken* auf. Dann verwendet man die *Ergebnisfenster-Optionen,* um nur diese beiden Faktoren auszuwählen:

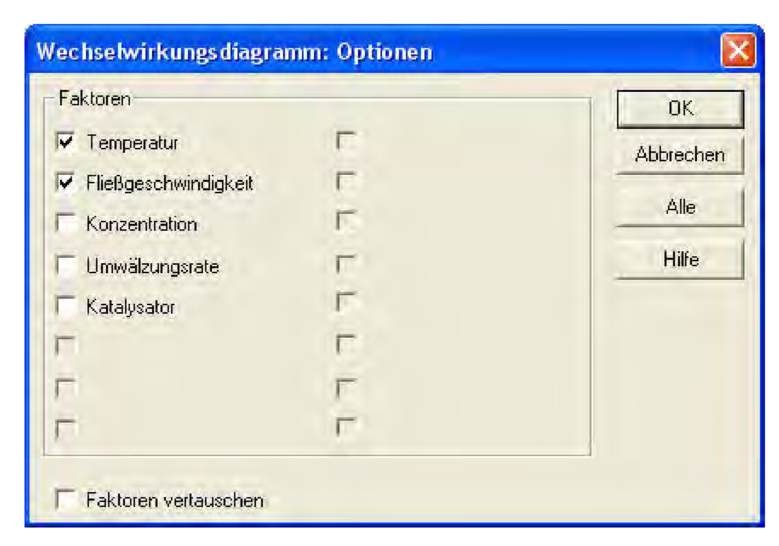


Abbildung 16.19: Ergebnisfenster-Optionen für das Wechselwirkungs-Diagramm

Das resultierende Diagramm zeigt den mittleren *Ertrag* für die sich ändernden *Temperaturen* für jede Stufe von *Fließgeschwindigkeit*:

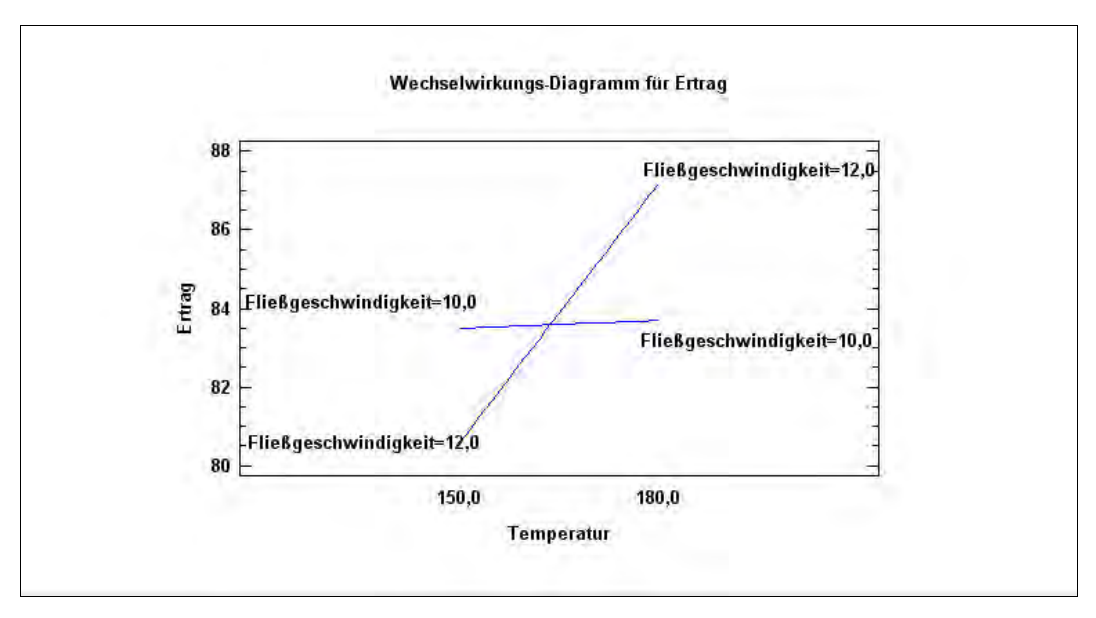


Abbildung 16.20: Wechselwirkungs-Diagramm für Fließgeschwindigkeit und Temperatur

Beachten Sie, dass die *Temperatur* bei der geringen *Fließgeschwindigkeit* – wenn überhaupt – nur einen kleinen Effekt hat. Bei der höheren *Fließgeschwindigkeit* ist *Temperatur* ein sehr wichtiger Faktor.

Bevor das statistische Modell verwendet wird, das hier zugrunde liegt, ist es wichtig, die nicht-signifikanten Faktoren zu entfernen. Um dies zu tun:

1. Drücken Sie die Schaltfläche *Analyse-Optionen* in der Analyse-Symbolleiste.

2. Drücken Sie die Schaltfläche *Entfernen* in der Dialogbox der *Analyse-Optionen*.

3. In der Dialogbox *Entfernen von Effekten: Optionen* klicken Sie doppelt auf jeden Effekt, den Sie ausschließen möchten, wodurch dieser von der Spalte *Einbeziehen* in die Spalte *Entfernen* verschoben wird:



Abbildung 16.21: Dialog-Box für Entfernen von Effekten

Nach folgenden Regeln sollte man Effekte ausschließen:

1. Jede nicht-signifikante Zweifach-Wechselwirkung ausschließen.

2. Jeden nicht-signifikanten Haupteffekt ausschließen, der nicht in einer signifikanten Zweifach-Wechselwirkung enthalten ist.

Für uns bedeutet das: Alle Effekte, die im Pareto-Diagramm nicht signifikant sind, werden ausgeschlossen, mit Ausnahme des Haupteffektes von B. Letzterer bleibt einbezogen, weil B in einer signifikanten Wechselwirkung mit Faktor A enthalten ist.

Nachdem die betreffenden Effekte ausgeschlossen sind, sollte das Pareto-Diagramm wie folgt aussehen:

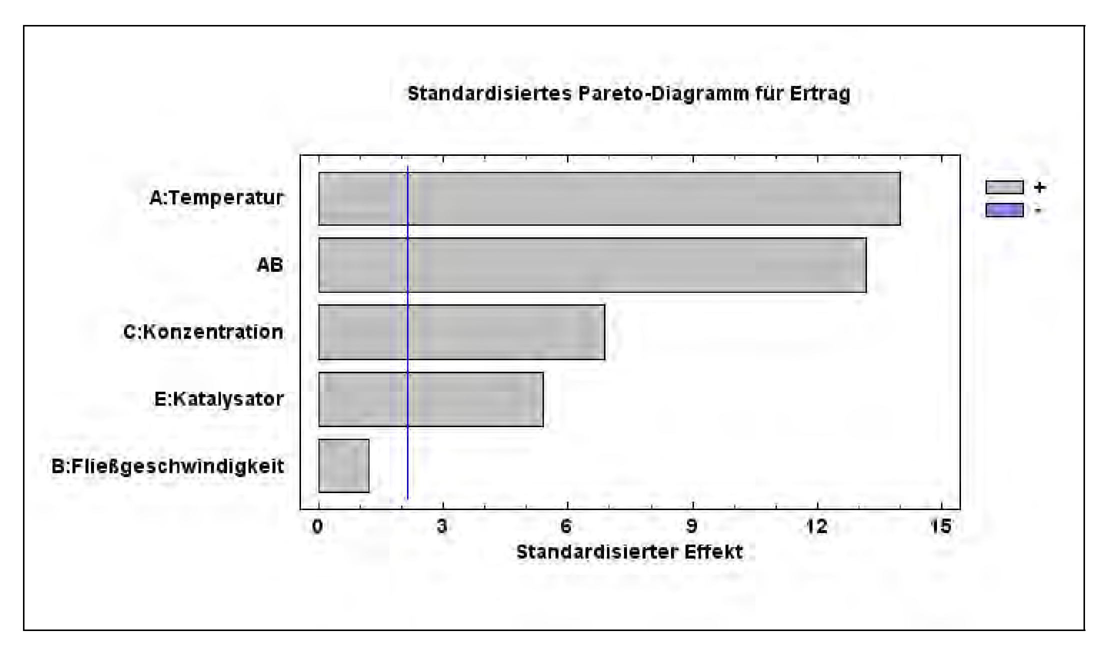


Abbildung 16.22: Standardisiertes Pareto-Diagramm nach dem Ausschließen von Effekten

Außer dem Haupteffekt von B sind alle verbleibenden Effekte statistisch signifikant. Das fertige Modell kann man sich anschauen, indem man die *Regressionskoeffizienten* in der Dialogbox *Tabellen und Grafiken* auswählt:

Regressionkoeffizienten für Ertrag

Koeffizient	*Schätzwert*
Konstante	250,074
A:Temperatur	-1,0595
B:Fließgeschwindigkeit	-17,4475
C:Konzentration	0,555417
E:Katalysator	2,6175
AB	0,106625

Der StatAdvisor

In diesem Fenster wird die Regressionsgleichung für die Anpassung an die Daten angezeigt. Die Gleichung des angepassten Modells ist die folgende:

Ertrag = 250,074 - 1,0595*Temperatur - 17,4475*Fließgeschwindigkeit + 0,555417*Konzentration + 2,6175*Katalysator + 0,106625 *Temperatur*Fließgeschwindigkeit

Abbildung 16.23: Angepasstes Regressionsmodell für Ertrag

Beachten Sie, dass das zugrunde liegende Modell die Form eines linearen Mehrfach-Regressionsmodells hat. Jeder verbleibende Haupteffekt ist selbst im Modell enthalten, die Zweifach-Wechselwirkung wird durch ein Kreuzprodukt aus *Temperatur* und *Fließgeschwindigkeit* repräsentiert.

Um das angepasste Modell vollständig verstehen zu können, stellt man es am besten grafisch dar. Wenn man die *Zielgrößen-Diagramme* in der Dialogbox *Tabellen und Grafiken* auswählt, können verschiedene Arten von Grafiken erzeugt werden. Standardmäßig erscheint ein drahtgitterförmiges Oberflächendiagramm:

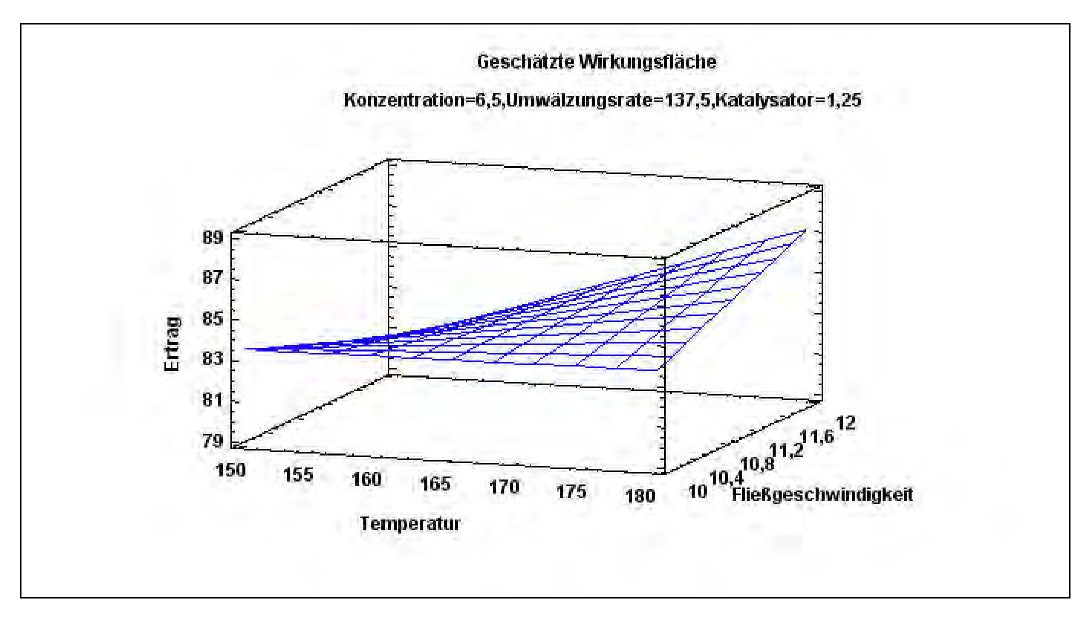


Abbildung 16.24: Zielgrößen-Diagramm

In diesem Diagramm repräsentiert die Höhe der Fläche den Modellwert von *Ertrag* im Versuchsraum von *Temperatur* und *Fließgeschwindigkeit*, wobei die anderen drei Faktoren auf ihren mittleren Werten konstant gehalten werden. Der höchste Ertrag wird bei der höchsten Temperatur und der höchsten Fließgeschwindigkeit erreicht.

Die Art der Grafik sowie die einbezogenen Faktoren können mittels *Ergebnisfenster-Optionen* geändert werden:

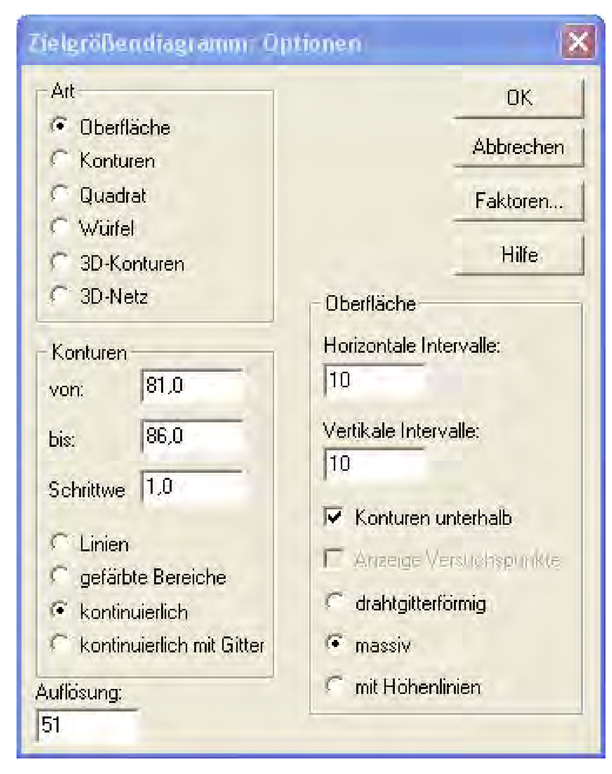


Abbildung 16.25: Ergebnisfenster-Optionen für das Zielgrößen-Diagramme

Die Grafikarten, die man erzeugen kann, sind die folgenden:

1. *Oberfläche*: zeichnet die angepasste Gleichung als eine dreidimensionale Fläche in Bezug auf jeweils zwei experimentelle Faktoren. Die Fläche kann wie ein *Drahtgitter* aussehen, *massiv* in einer Farbe erscheinen oder *Höhenlinien* für die Zielgröße anzeigen. *Konturen unterhalb* bezieht ein Konturen-Diagramm in die Grundfläche der Grafik ein.

2. *Konturen*: erzeugt ein zweidimensionales Konturen-Diagramm in Bezug auf jeweils zwei experimentelle Faktoren. Die Konturen können als (Höhen-) *Linien* wie auf einer topografischen Karte, als *gefärbte Bereiche* oder unter Verwendung einer *kontinuierlichen* Farbpalette dargestellt werden.

3. *Quadrat*: zeichnet den Versuchsraum für jeweils zwei experimentelle Faktoren und zeigt den Zielgrößenwert in jeder Ecke des Quadrates an.

4. *Würfel*: zeichnet den Versuchsraum für jeweils drei experimentelle Faktoren und zeigt den Zielgrößenwert in jeder Ecke des Würfels an. Für dieses Diagramm müssen Sie

zunächst die Schaltfläche *Faktoren* drücken und einen dritten Faktor auswählen.

5. *3D-Konturen*: zeichnet die Konturen der Zielgröße in bezug auf 3 experimentelle Faktoren gleichzeitig.

6. *3D-Netz*: zeichnet ein Netzdiagramm, das den Wert der Zielgröße innerhalb eines dreidimensionalen Versuchsraumes darstellt.

Die Schaltfläche *Faktoren* dient zur Auswahl der Faktoren, die an die Achsen der Diagramme gezeichnet werden sollen, und der Festlegung der Werte für die anderen Faktoren, die nicht dargestellt werden.

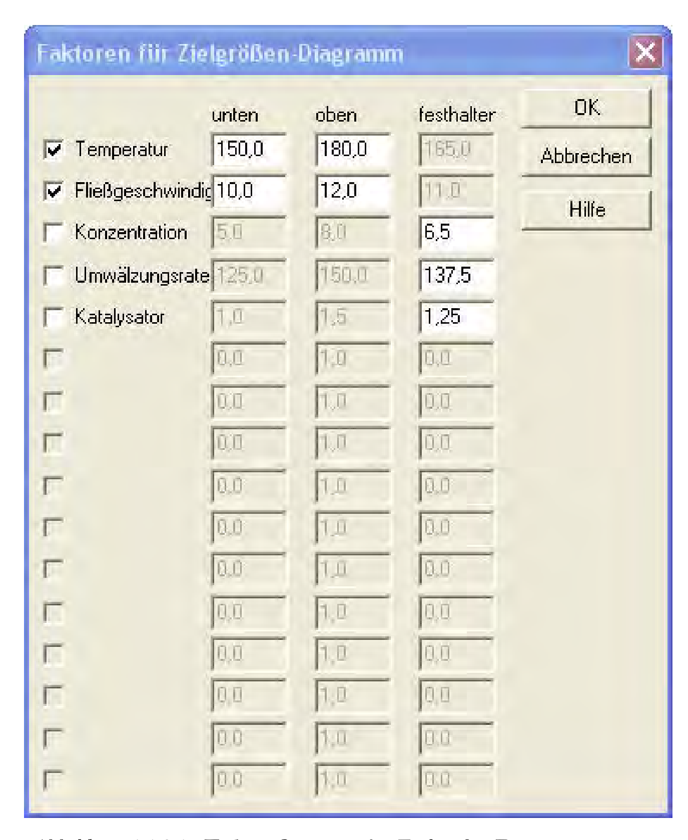


Abbildung 16.26: Faktor-Optionen für Zielgrößen-Diagramme

Um das nachfolgende Diagramm zu erzeugen, wurde das Gruppenfeld *Konturen* auf *kontinuierlich*, *Fläche* auf *massiv* sowie *Konturen unterhalb* gesetzt, und die Höhenlinien wurden im Bereich von 81 bis 86 mit einer Schrittweite 1 festgelegt:

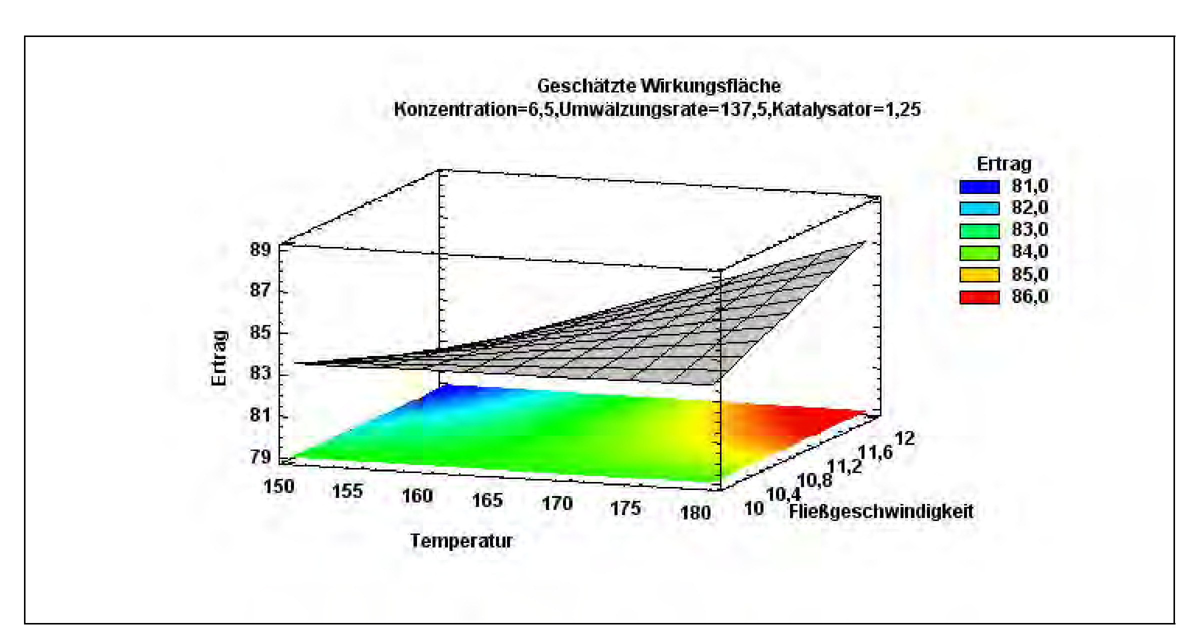


Abbildung 16.27: Zielgrößen-Diagramm mit unten eingeblendeten Konturen

Die gleiche Grafik kann als Konturen- anstatt als Flächen-Diagramm angezeigt werden:

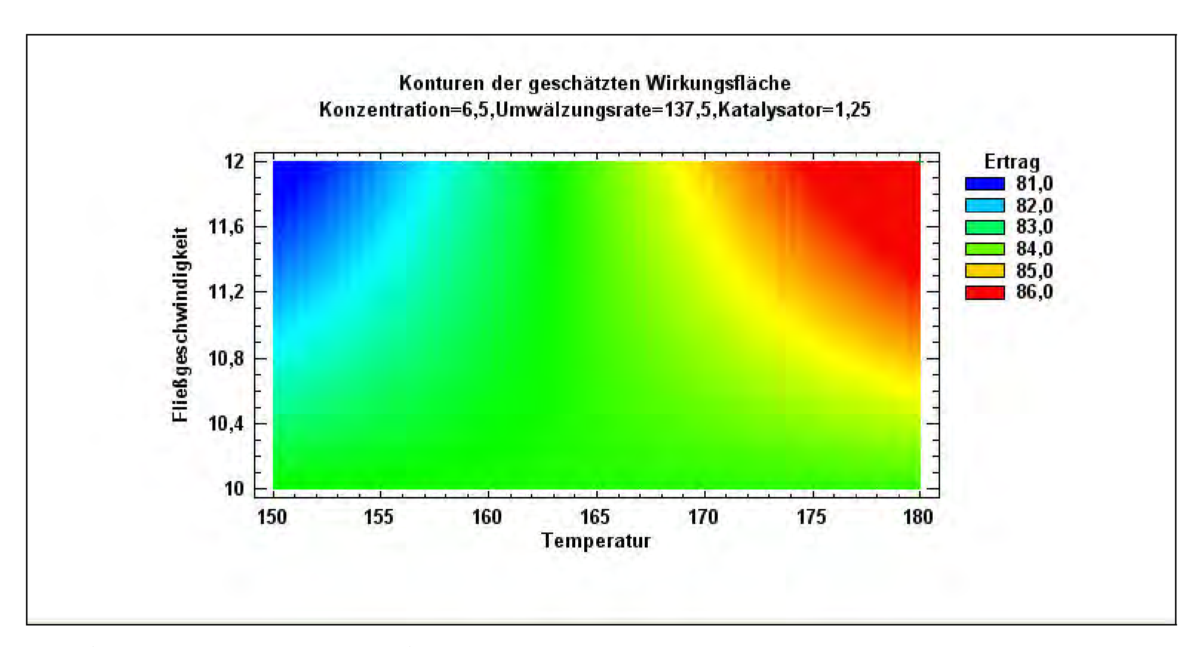


Abbildung 16.28: Konturen-Diagramm der Zielgröße

Die höchsten Werte von *Ertrag* findet man in der rechten oberen Ecke.

Die zweite während der Experimente gemessene Zielgröße war *Bruchfestigkeit*. Deren Analysefenster zeigt das folgende Pareto-Diagramm:

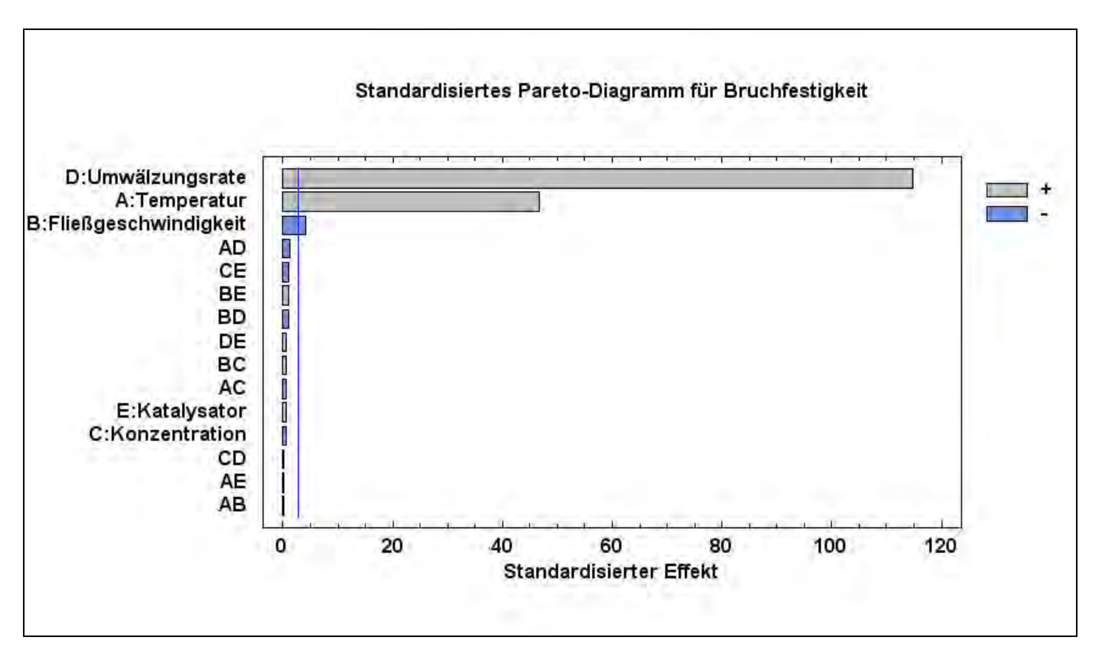


Abbildung 16.29: Standardisiertes Pareto-Diagramm für Festigkeit

Nach der Entfernung der nichtsignifikanten Effekte lautet das angepasste Modell:

Bruchfestigkeit = -317,288 + 1,02083*Temperatur - 1,3125*Fließgeschwindigkeit + 3,005*Umwälzungsrate

Beachten Sie, dass die *Umwälzungsrate* die *Bruchfestigkeit* beeinflusst, aber nicht den *Ertrag*. Das Konturendiagramm für die beiden stärksten Einflussfaktoren sieht folgendermaßen aus:

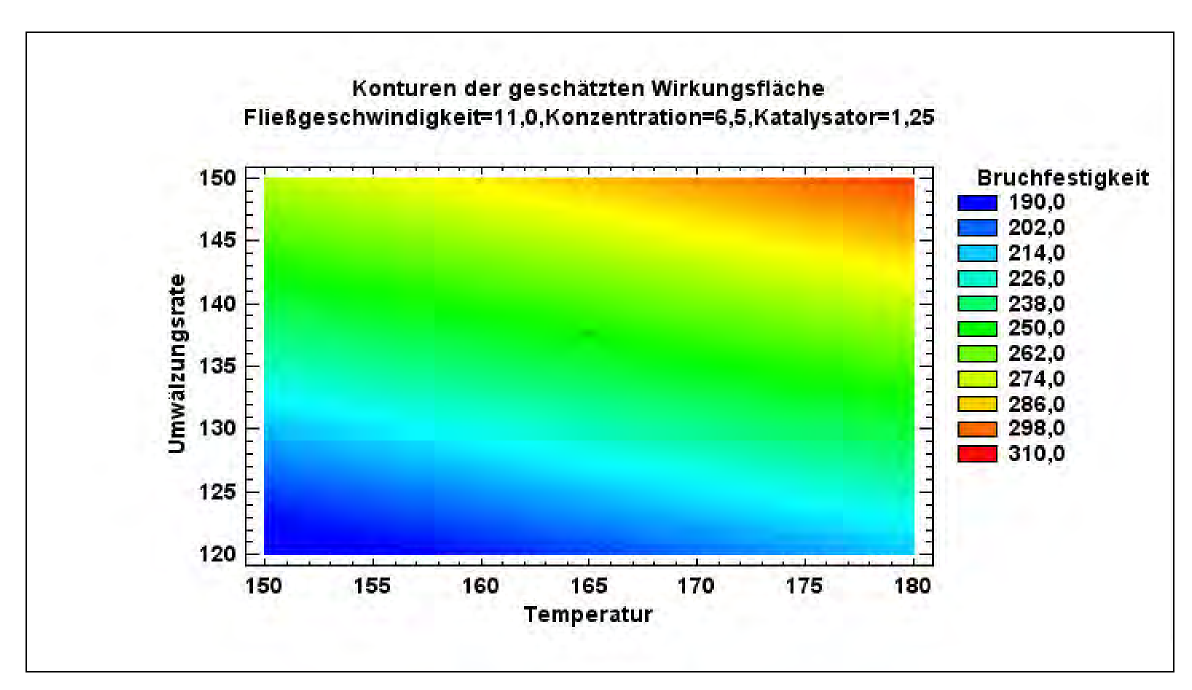


Abbildung 16.30: Konturendiagramm für Festigkeit

Stufe 9: Zielgrößen optimieren

Nachdem für jede Zielgröße ein statistisches Modell gebildet wurde, können nun die optimalen Faktoreinstellungen bestimmt werden. Zur Erinnerung: Das Ziel des Experiments war es, den *Ertrag* zu maximieren, während die *Bruchfestigkeit* dem Wert 250 psi möglichst nahe kommen soll. Nach dem Drücken der Schaltfläche *Stufe 9* wird folgende Dialogbox angezeigt:

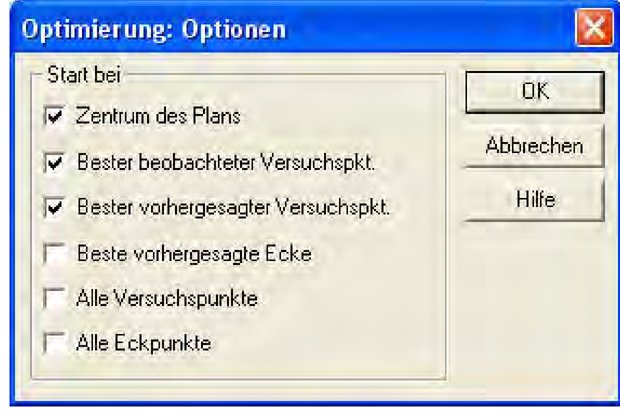


Abbildung 16.31. Dialogbox mit den Optionen für die Optimierung der Zielgrößen

Da das Programm eine numerisches Suchverfahren anwendet, um den besten Punkt innerhalb des Versuchsraumes zu finden, ist es eine gute Idee, die Suche von verschiedenen Punkten aus zu starten, um ein lokales Optimum zu finden.
Drücken Sie *OK*, um die Suche zu beginnen. Nach einem kurzen Moment wird folgende Meldung angezeigt:



Abbildung 16.32: Meldung, wenn die Optimierung abgeschlossen ist

Gleichzeitig werden folgende Ergebnisse in das Hauptfenster des DoE-Assistenten eingefügt:

Stufe 9: Optimierung der Zielgrößen

Zielgrößenwerte im Optimum

Zielgr.	*Vorhersage*	*Untere 95,0% Grenze*	*Obere 95,0% Grenze*	*Erwünschtheit*
Ertrag	88,6736	88,0469	89,3002	0,867359
Bruchfestigkeit	250,0	249,047	250,953	0,999999

Gesamt-Erwünschtheit = 0,948035

Faktoreinstellungen im Optimum

Faktor	*Einstellung*
Temperatur	180,0
Fließgeschwindigkeit	12,0
Konzentration	8,0
Umwälzungsrate	132,874
Katalysator	1,49999

Abbildung 16.33: Zusammenfassung der Optimierung im Hauptfenster des DoE-Assistenten

Bei den angegebenen Faktoreinstellungen wird der *Ertrag* mit 88,67 Gramm geschätzt und die *Bruchfestigkeit* mit 250 psi. Der resultierende *Ertrag* hat einen "Erwünschtheits"- Quotienten von 0,867, das sind 86,7% des Weges im festgelegten Bereich zwischen 80 und 90 Gramm. *Festigkeit* hat einen Erwünschtheits-Quotienten von 1, da hier genau der Zielwert getroffen wird. Die Gesamt-Erwünschtheit liegt bei 0,948. Dieser Wert wird berechnet, indem die Erwünschtheit für jede einzelne Zielgröße mit ihrer *Bedeutung* potenziert, aus diesen das Produkt gebildet und dann mit 1 durch die Summe der Bedeutungen potenziert wird. Das Ergebnis ist eine Zahl zwischen 0 und 1, wobei der Zielgröße mit einer höheren Bedeutung eine höhere Gewichtung gegeben wird. Nach dem Drücken der Schaltfläche *Tabellen und Grafiken* können zwei weitere Diagramme

erzeugt werden. Das *Überlagerungsdiagramm* zeigt die übereinandergelegten Konturendiagramme der beiden Zielgrößen:

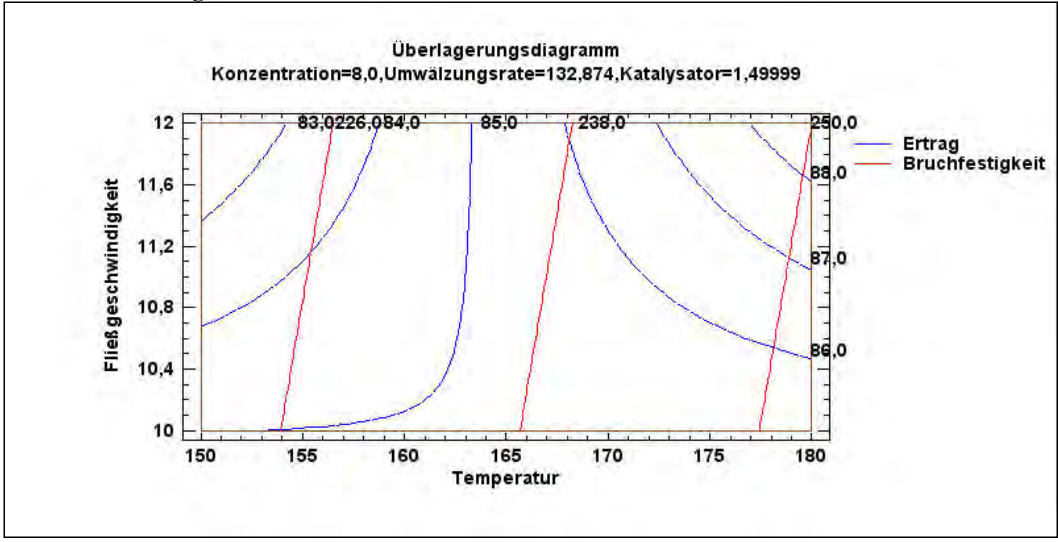


Abbildung 16.34:. Übereinandergelegte Konturendiagramme für die beiden Zielgrößen

Der optimale Punkt befindet sich in der oberen rechten Ecke, wo der *Ertrag* entlang der Linie für *Festigkeit*=250 maximiert wird. Das *Diagramm der Wunschfunktion* kann zur Anzeige der Gesamt-Erwünschtheit über jeweils zwei oder drei Faktoren verwendet werden. Bei Auswahl eines 3D-Netzdiagramms sieht das folgendermaßen aus:

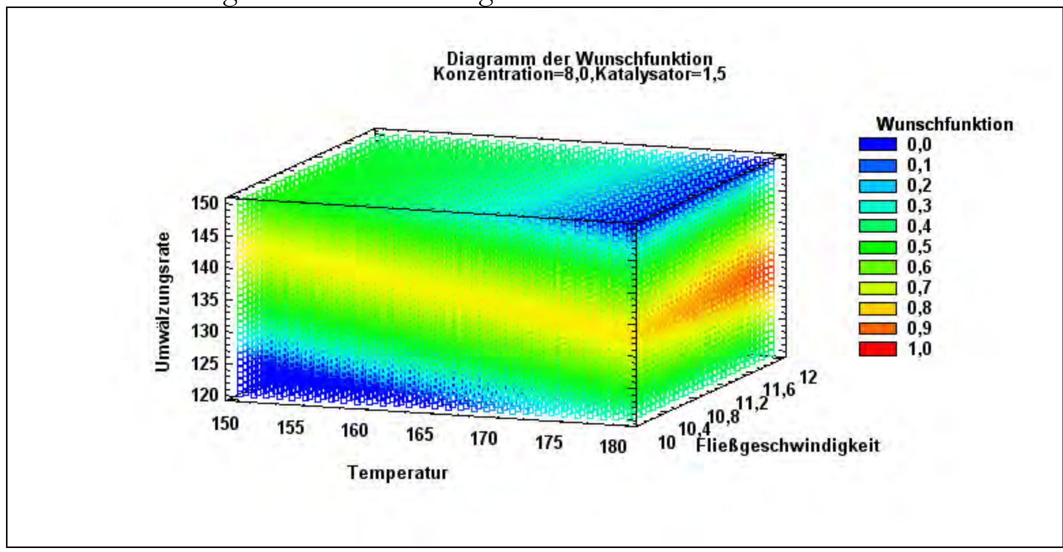


Abbildung 16.35: 3D-Netzdiagramm der Gesamtwunschfunktion

Die beste Lage wird in Rot angezeigt, wo sowohl *Temperatur* als auch *Fließgeschwindigkeit* hoch sind, während *Umwälzungsrate* auf einem mittleren Wert liegt.

Stufe 10: Ergebnisse speichern

Um die Ergebnisse der Analyse und der Optimierung in eine StatFolio-Datei zu speichern, drücken Sie die Schaltfläche *Stufe 10*:

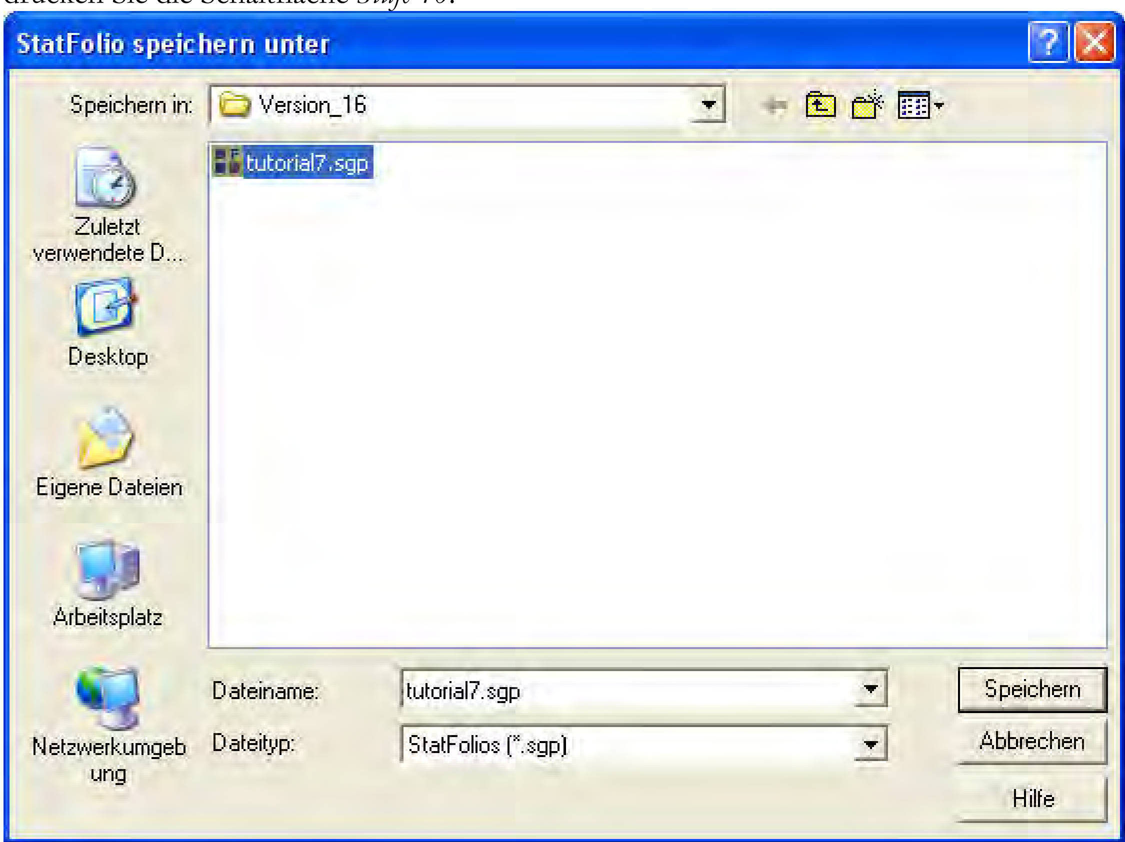


Abbildung 16.36: Dialogbox zum Speichern der Ergebnisse

16.3 Fortsetzen der Experimente

Falls weiteres Experimentieren gewünscht ist, kann STATGRAPHICS Centurion XVI dabei helfen, den Versuchsplan zu ergänzen oder Punkte entlang des Pfades des steilsten Anstiegs zu generieren.

Stufe 11: Versuchsplan ergänzen

Nach Drücken der Schaltfläche *Stufe 11* können Sie dem aktuellen Versuchsplan weitere Versuche hinzufügen. Folgende Dialogbox wird angezeigt:

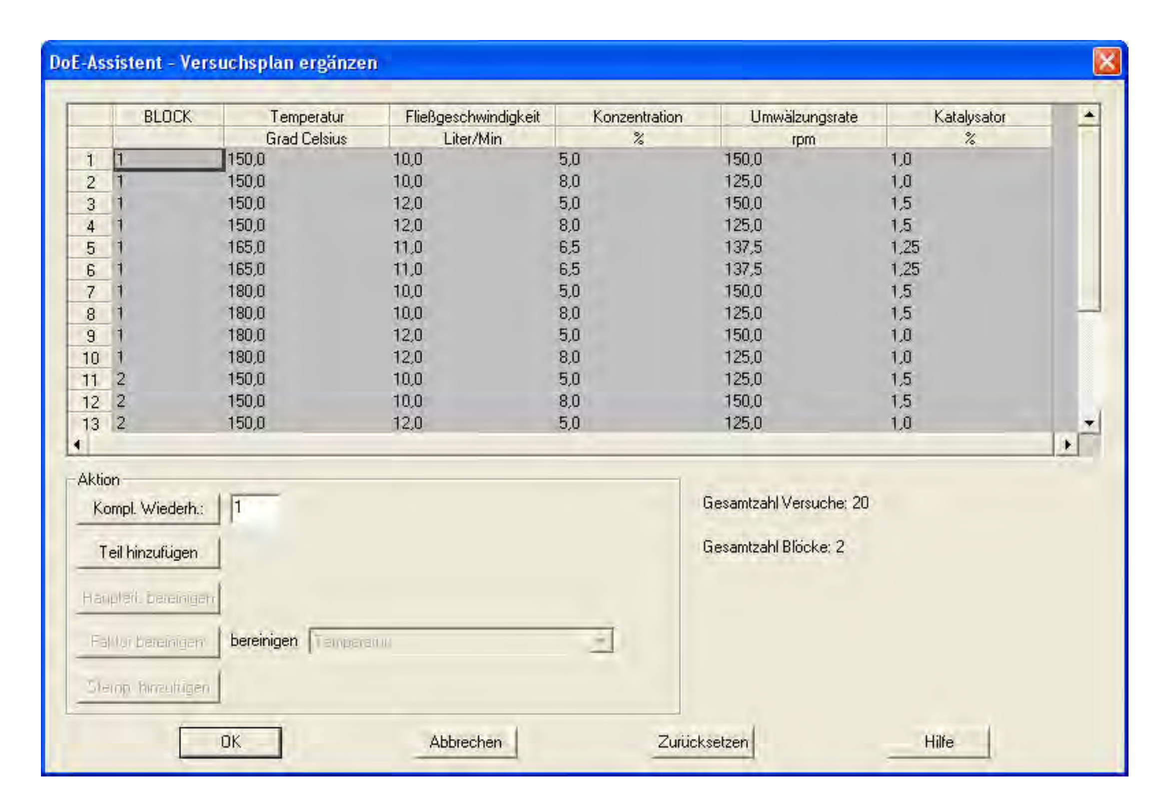


Abbildung 16.37: Dialogbox Versuchsplan ergänzen

Es sind drei Optionen verfügbar:

1. *Komplette Wiederholung:* fügt dem Versuchsplan weitere 20 Versuche hinzu, die identisch mit den ersten 20 sind. Dadurch erhält man mehr Freiheitsgrade für die Schätzung des Versuchsfehlers.

2. *Zusammenschieben:* entfernt einen festgelegten Faktor aus dem Plan und den Auswertungen.

3. *Teil hinzufügen:* fügt weitere 20 Versuche hinzu, um aus dem aktuellen Plan einen Vollfaktorplan zu erzeugen.

Stufe 12: Extrapolieren

Sie können auch Punkte entlang des *Pfades des steilsten Anstiegs* generieren, um zu versuchen, die Bereiche mit höherem Ertrag schneller zu erreichen. Der Pfad des steilsten Anstieges ist der Weg, der an einem festgelegten Punkt des Versuchsraumes beginnt und sich in Richtung der größten Änderung der geschätzten Zielgröße bei kleinsten Faktor-Veränderungen bewegt. Wenn man diesem Pfad folgt, kann man sehr effektiv und schnell deutliche Verbesserungen erreichen. Nach Drücken der Schaltfläche *Stufe 12* wird folgende Dialogbox angezeigt:

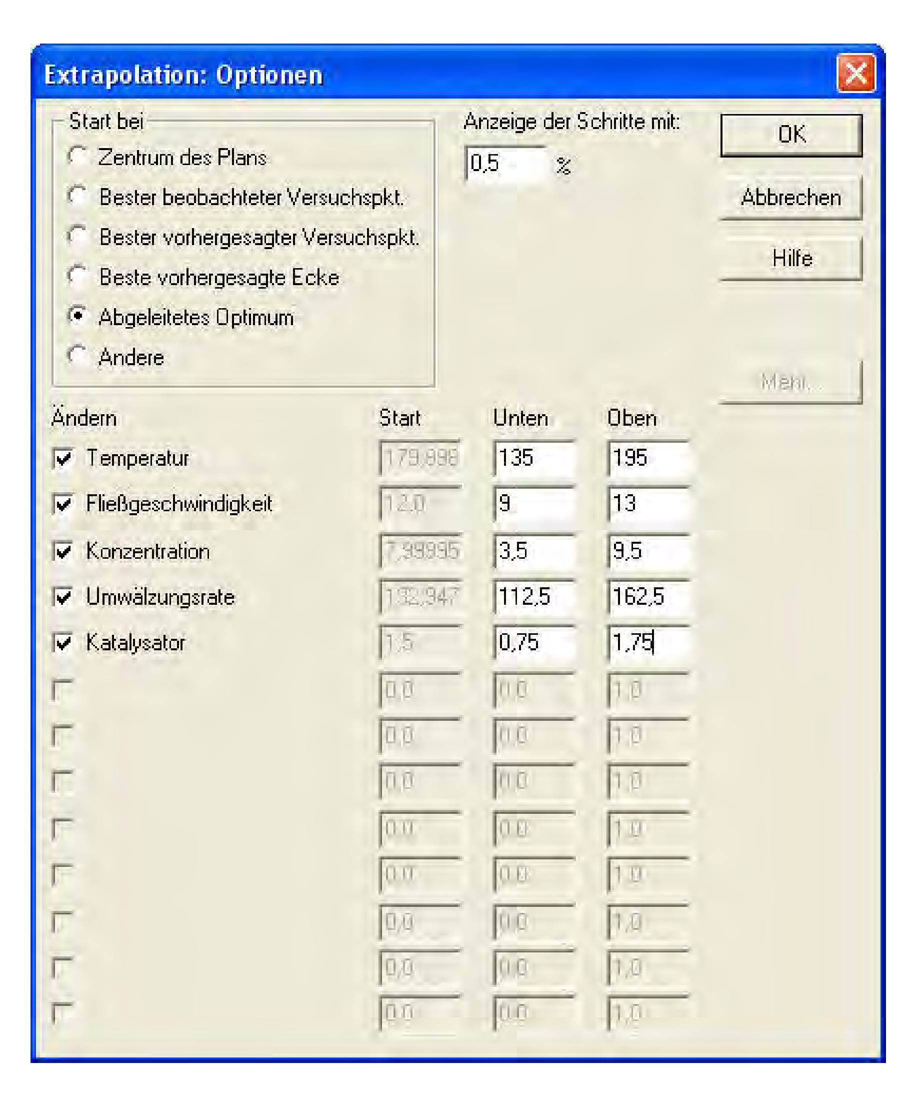


Abbildung 16.38: Dialogbox für Extrapolieren (Pfad des steilsten Anstiegs)

Die Einstellungen in dieser Dialogbox veranlassen das Programm, beim erreichten Optimum zu beginnen und die 5 Faktoren zwischen *Unten* und *Oben* variieren zu lassen, was die Breite Versuchsraumes in jeder Dimension verdoppelt. Es sollen Faktorkombinationen angezeigt werden, sobald sich die geschätzte Erwünschtheit mindestens um 0,5% ändert. Nachdem *OK* gedrückt wurde, wird folgende Tabelle dem Fester des DoE-Assistenten hinzugefügt:

Stufe 12: Modell-Extrapolation

Extrapolierte Zielgrößenwerte

Schritt	Erwünschtheit	Ertrag	Bruchfestigkeit
0	0,948022	88,6736	249,999
1	0,953451	88,8107	249,985
2	0,958935	88,9483	249,979
3	0,96487	89,094	250,012
4	0,97038	89,2326	250,01
5	0,975828	89,376	250,025
6	0,981255	89,519	250,035
7	0,987743	89,6839	250,023
8	0,993077	89,8332	250,051
9	0,99884	89,9783	249,972
10	0,999491	90,0045	249,959

Faktoreinstellungen für die Extrapolation

Schritt	Temperatur	Fließgeschwindigkeit	Konzentration	Umwälzungsrate	Katalysator
0	180,0	12,0	8,0	132,874	1,49999
1	180,051	12,05	8,04282	132,874	1,5056
2	180,109	12,1	8,0836	132,874	1,51093
3	180,206	12,15	8,12264	132,874	1,51605
4	180,268	12,2	8,16179	132,874	1,52117
5	180,347	12,25	8,20118	132,874	1,52633
6	180,421	12,3	8,24027	132,874	1,53145
7	180,486	12,36	8,28549	132,874	1,53737
8	180,578	12,41	8,32475	132,874	1,5425
9	180,578	12,47	8,36554	132,874	1,54784
10	180,578	12,48	8,37467	132,874	1,54904

Abbildung 16.39: Zusammengefasste Extrapolation im Hauptfenster des DoE-Assistenten

Es wird geschätzt, dass der Ertrag auf seinen Zielwert von 90 Gramm gesteigert werden kann, während die Festigkeit bei 250 gehalten wird, indem die Temperatur auf 180,6 Grad, die Fließgeschwindigkeit auf 12,48 Liter/Min, die Konzentration auf 8,37% und der Katalysator auf 1,55% erhöht werden.

Da es sich um eine Extrapolation des angepassten statistischen Modells außerhalb des ursprünglichen Versuchsraumes handelt, sind Bestätigungsversuche zur Überprüfung dieses Ergebnisses erforderlich.

Literaturhinweise

In den folgenden Büchern finden Sie ausgezeichnete weiterführende Informationen zu den statistischen Techniken, die in diesem Handbuch beschrieben werden :

Statistische Grundlagen: Applied Statistics and Probability for Engineers, 4th edition, von Douglas C. Montgomery und George C. Runger (2006). John Wiley and Sons, New York.

Varianzanalyse: Applied Linear Statistical Models, 5th edition, von Michael H. Kutner, Christopher J. Nachtsheim und John Neter (2004). McGraw Hill.

Regressionsanalyse: Applied Linear Regression, 3rd edition, von Sanford Weisberg (2005). John Wiley and Sons, New York.

Statistische Prozesssteuerung: Introduction to Statistical Quality Control, 6th edition, von Douglas C. Montgomery (2008). John Wiley and Sons, New York.

Versuchsplanung: Statistics for Experimenters: Design, Innovation and Discovery, 2nd edition, von George E. P. Box, William G. Hunter und J. Stuart Hunter (2005). John Wiley and Sons, New York.

Daten

93cars.sgd

Diese Daten sind ein Download vom Journal of Statistical Education (JSE) Data Archive. Sie wurden von Robin Lock aus dem Mathematics Department at St. Lawrence University zusammengestellt und werden mit seiner Erlaubnis verwendet. Ein Artikel in Zusammenhang mit diesen Daten erschien im *Journal of Statistics Education*, Volume 1, Number 1 (July 1993).

bodytemp.sgd

Diese Daten sind ebenfalls ein Download vom Journal of Statistical Education (JSE) Data Archive. Sie wurden von Allen Shoemaker aus dem Psychology Department at Calvin College zusammengestellt und werden mit seiner Erlaubnis verwendet. Die Daten wurden aus einem Artikel im *Journal of the American Medical Association* (1992, vol. 268, pp. 1578-1580) mit dem Titel "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" von P. A. Mackowiak, S. S. Wasserman, und M. M. Levine abgeleitet. Ein Artikel in Zusammenhang mit diesen Daten erschien im *Journal of Statistics Education*, Volume 4, Number 2 (July 1996).

Internetseite vom Journal of Statistical Education (JSE) Data Archive:

http://www.amstat.org/publications/jse/jse_data_archive.htm

Stichwortverzeichnis